Understanding retinal diseases with genotypic and transcriptomic data analysis

Satyasai Aravind Prasad MANDA
ORCID: 0000-0002-5497-1705

Master of Philosophy
September 2021
Faculty of Medicine, Dentistry and Health Sciences
University of Melbourne

A thesis submitted in total fulfilment of the requirements for the degree of Master of Philosophy

Supervisors:
Prof. Melanie Bahlo & Dr Brendan Ansell
Department of Medical Biology, WEHI, University of Melbourne.

Declaration

This is to certify that:

- this thesis comprises only my original work towards the Master of Philosophy except where indicated in the Preface;
- due acknowledgement has been made in the text to all other material used;
- this thesis is fewer than 50,000 words in length, exclusive of tables, bibliographies and appendices.

Satyasai Aravind Prasad Manda
Faculty of Medicine, Dentistry and Health Sciences,
The University of Melbourne.
Bahlo Laboratory, WEHI.

Preface

The research work in this thesis was undertaken in the laboratory of Prof. Melanie Bahlo, Population Health and Immunity Division, WEHI / Department of Medical Biology, University of Melbourne.

The following information includes the contributions and publication status for each chapter. Publications appear in part in this thesis. My contribution to these publications is approximately 10%, and the details are included in Chapter 4. I have provided proper citations for all the material reproduced or used in this thesis.

I acknowledge the contributions of the following individuals to the work in this thesis:

Chapter 1: Introduction and Literature Review. This chapter introduces various techniques and previous studies relevant to my research work. It comprises a summary of the elaborate literature review and is an original work, with advisory comments as well as assistance in editing, reviewing and grammar from Prof. Melanie Bahlo and Dr Brendan Ansell.

Chapter 2: Description and data processing of a novel retinal eQTL dataset. This is an original unpublished work in which I have performed the pre-processing and quality control steps for our data.
The contributions to this work are as follows:

- Prof. Melanie Bahlo and Dr Brendan Ansell conceived and directed this work.
- I performed the data processing and quality control analysis of the genotype and RNA-seq data.
- I interpreted and created visualisations from the data with inputs from Prof. Melanie Bahlo and Dr Brendan Ansell.

- I wrote this chapter with assistance in editing, reviewing and grammar from Prof. Melanie Bahlo and Dr Brendan Ansell.
- The genotype data was provided generously by our collaborators Prof. Sandro Banfi and Dr Diego Di Bernardo from TIGEM, Italy.

Chapter 3: eQTL analysis of a novel retinal dataset. This is an original unpublished work in which I integrated the genotype and RNA-seq data after quality control to identify eQTLs in the retina and performed a comparative analysis with published studies.

The contributions to this work are as follows:

- Prof. Melanie Bahlo and Dr Brendan Ansell conceived and directed this work.
- I performed the identification of retinal eQTLs and the associated data analysis including troubleshooting and exploring the eQTL analysis workflow. The ideas based on statistical genetics for data exploration were provided by Prof. Melanie Bahlo and Dr Brendan Ansell.
- I interpreted and visualised the data with inputs from Prof. Melanie Bahlo and Dr Brendan Ansell.
- I wrote this chapter with assistance in editing, reviewing and grammar from Prof. Melanie Bahlo and Dr Brendan Ansell.

Chapter 4: Regulation of retinal gene expression in MacTel. This is an original work that forms part of a publication in a reputed peer-reviewed journal.

R. Bonelli, V. Jackson *et al.*, **'Identification of genetic factors influencing metabolic dysregulation and retinal support for MacTel, a retinal disorder'**, *Commun. Biol.*, vol. 4, no. 1, Mar. 2021, doi: 10.1038/s42003-021-01788-w.

The contributions for the publication work are as follows:

- Prof. Melanie Bahlo and Dr Brendan Ansell conceptualised, supervised, written and reviewed this work.
- Dr Roberto Bonelli and Dr Victoria Jackson have conceptualised, investigated, performed the data analysis, visualised, written and reviewed this work.

- My contributions to this publication are 10% overall with formal data analysis and investigation related to eQTLs in MacTel.

This chapter is based on the research work performed by me for this publication.

- I wrote this chapter with assistance in editing, reviewing and grammar from Prof. Melanie Bahlo and Dr Brendan Ansell.

Chapter 5: Discussion. This is an original work in which I have discussed and summarised the importance of the research work performed as a part of this thesis. I wrote this chapter with assistance in editing, reviewing and grammar from Prof. Melanie Bahlo and Dr Brendan Ansell.

Acknowledgements

All along my journey towards Masters' completion, I have received commendable support and assistance and would like to acknowledge everyone for their encouragement.

First and foremost, I am immensely thankful to my principal supervisor, Melanie Bahlo, for her excellent guidance and support during this course. She has been a great supervisor and a fantastic motivator that made me feel excited scientifically, after discussions with her. I appreciate her honest feedback which made me learn and perform better. She has been a great deal of support during my difficulties in various forms. I am well benefited from her extensive professional experience and versatile leadership to improve myself.

I am profoundly grateful to my co-supervisor, Brendan Ansell for his invaluable support and for being an excellent mentor throughout my candidature. He has been very helpful to gain valuable skills in various aspects including data visualisation and interpretation in R and scientific communication that are important for my professional development. I truly appreciate his prompt guidance, feedback and being generous with his time. It has been a fantastic and outstanding learning experience working with him.

My sincere thanks to my friendly advisory committee, Vanessa Bryant, Erica Fletcher, Peter Hickey, Robyn Guymer, Yunshun Chen, for their constructive ideas and encouragement. I appreciate their time during the meetings.

My sincere thanks to my collaborators from TIGEM, Italy for generously providing us with the genotype data that made the analysis possible for my thesis.

Many thanks to all the members of Bahlo lab for their co-operation and friendliness to learn from them. They have been an excellent group to work with and chat together. Special thanks to Vicki Jackson, Jacob Munro, Roberto Bonelli, Samaneh Farashi for their help and advice on various technical aspects and coding. I thank the PhD students from our group Jiru Han, Karen Oliver, Erandee Robertson for their help and support.

My sincere thanks to Keely Bumsted O'Brien for her mentoring sessions and the Scientific Education Office of WEHI for offering valuable upskilling and career planning workshops.

I would also thank Natalie Senzo and Sarah Miller for organizing various meetings during my candidature.

My research would not have been possible without the fee-offset scholarship support from the University of Melbourne and stipend support from the MacTel project and I am grateful to them.

Thanks to my friends, in India and Australia, for their support and cheering me up and lending their ears and time when needed.

"To my mother in her loving memory". I am ever thankful to my parents for always supporting me in my endeavours and boosting my courage at all times.

Finally, to my partner, Dedeepya Vadali, for her patience, moral support and understanding all the time.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AMD	Age-related macular degeneration
ANOVA	Analysis of variance
BED	Browser Extensible Data format
CPM	Counts per million
CV	Coefficient of variation
DNA	Deoxyribonucleic acid
DNA-seq	DNA sequencing; whole genome or whole exome sequencing technology
EBI	European Bioinformatics Institute
eGene	A gene containing eQTLs
eQTL	Expression quantitative trait locus
eSNP	An eQTL SNP; SNP associated with the expression of an eGene
EUR	European
EyeGEx	Eye Genotype Expression
FDR	False discovery rate
FUMA-GWAS	Functional Mapping and Annotation of Genome-Wide Association Studies server
GEO	Gene expression omnibus
GERA	Genetic Epidemiology Research on Aging
GTEx	Genotype-Tissue Expression
GTF	Gene transfer format
GWAS	Genome wide association studies
GW	Genome wide
h^2	h-squared; narrow sense heritability estimate
HDL	High-density lipoprotein
HGP	Human genome project
HRC	Haplotype reference consortium
HWE	Hardy-Weinberg Equilibrium
IAMDGC	Age-related Macular Degeneration Genomics Consortium
IBD	Identity By Descent

indel	Insertion or deletion
kb	Kilo base pairs
LCLs	Lymphoblastoid cell lines
LD	Linkage disequilibrium
LMM	Linear Mixed Model
lncRNAs	Long non-coding RNAs
LRAT	Lecithin retinol acyltransferase
MacTel	Macular telangiectasia type II
MAF	Minor allele frequency
Mb	Mega base pairs
miR	Micro RNA
mRNA	Messenger RNA
mtDNA	Mitochondrial DNA
ncRNA	Non-coding RNA
OR	Odds ratio
PC	Principal component
PCA	Principal component analysis
PP	Posterior probability
QC	Quality control
QTL	Quantitative trait locus
r^2	r-squared; quantification of linkage disequilibrium
RNA	Ribonucleic acid
RNA-seq	RNA sequencing technology
RPE	Retinal pigment epithelium
RPKM	Reads Per Kilobase of transcript per Million reads mapped
SAIGE	Scalable and Accurate Implementation of GEneralized mixed model
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variation
SPT	Serine palmitoyltransferase
sQTL	Splice quantitative trait locus
TGP	Thousand genomes project

TIGEM	Telethon Institute of Genetics and Medicine
ToPMed	Trans-Omics for Precision Medicine
TSS	Transcription start site
TWAS	Transcriptome-wide association studies
VCF	Variant calling format

URLs

ClinVar	www.ncbi.nlm.nih.gov/clinvar
coloc	CRAN.R-project.org/package=coloc
dbGaP	www.ncbi.nlm.nih.gov/gap
dbSNP	www.ncbi.nlm.nih.gov/snp
EBI ArrayExpress	www.ebi.ac.uk/arrayexpress/
Ensembl	www.ensembl.org/
FUMA-GWAS	fuma.ctglab.nl/
gnomAD	gnomad.broadinstitute.org
GTEx	www.gtexportal.org
GWAS catalog	www.ebi.ac.uk/gwas/
Limma	bioconductor.org/packages/limma/
Michigan Imputation Server	imputationserver.sph.umich.edu
PLINK	www.cog-genomics.org/plink/1.9/
powerEQTL	bwhbioinfo.shinyapps.io/powerEQTL/
QTLtools	qtltools.github.io/qtltools/
Rsubread	bioconductor.org/packages/Rsubread
STAR	github.com/alexdobin/STAR
TopMed	topmed.nhlbi.nih.gov/

ABSTRACT

The retina is light-sensitive eye tissue responsible for vision, but little is known about the genetic regulation of retinal gene expression. Investigating key drivers of gene regulation in the retina in healthy and diseased individuals remains a fundamental challenge in macular degeneration research, especially given the difficulty of accessing human retinal tissue. Deciphering the effects of genetic variation on retinal gene expression will underpin the development of novel treatment avenues for otherwise untreatable diseases causing blindness. A method to investigate these further focuses on the effects of genetic variants on gene expression levels derived from transcriptomic data. This type of 'omics analysis, known as expression quantitative trait (eQTL) analysis integrates genotype and gene-expression data.

The genotyping data for this thesis was generated in collaboration with scientists from the TIGEM, Italy, who first assembled the retinal transcriptome. We aimed to identify the genetic variants that modulate gene expression using a cohort of 41 individual donors of healthy retinal tissue. We performed retinal eQTL analysis using this independent cohort and compared our results with recently published retinal eQTL studies. After observing a weak eQTL signal potentially due to the small sample size, we explored potential strategies to mitigate the multiple testing burden so as to improve statistical power. To this end, we performed eQTL power analyses and limited both the set of variants and genes under consideration by thresholding on allele frequency and gene transcriptional abundance as well as disease relevance. Further, eQTL analysis was used to interpret the genetics of Macular Telangiectasia II, a blinding retinal degenerative disease. This included genome-wide and targeted interrogation of the signals from the largest genome-wide association study to date for this disease.

Chapter 1: Introduction and Literature review

Introduction

1.1 Physiology and anatomy of the eye

The human eye is a complex structure responsible for the sense of vision. The proper functioning of the eye relies on its ability to transduce light into electrical impulses by absorbing and processing light energy from the environment [1]. Human visual capacity is the output of continuous orchestrated functioning of different anatomical structures of the eye [2].

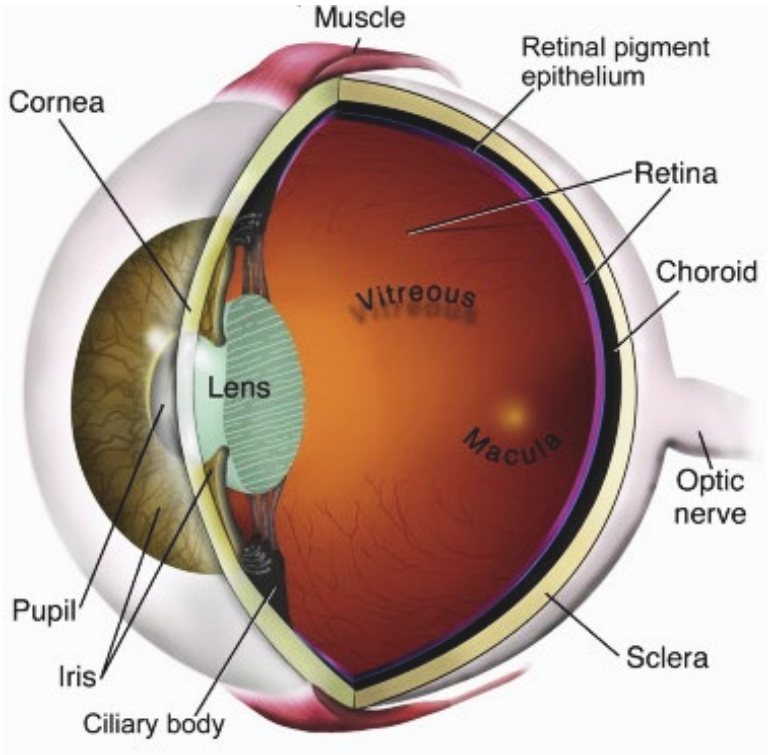

Figure 1.1: Pictorial representation of the physiological structure of the eye. Source: [3]

The major structures (Figure 1.1) of the human eye relevant for this thesis include a) the cornea: a clear and curved layer forming the front of the eye through which the light

enters, b) the iris: the coloured part of the outer eye behind the cornea controlling the amount of light that enters the eye through the pupil, an aperture at its centre, c) the lens: located directly behind the iris is the lens, a clear, crystalline disc-like structure with flexibility to change its thickness, with the help of ciliary muscles, to properly focus on an object depending on its distance from the eye, so-called accommodation; d) the retina: a layer of photosensitive cells covering the entire back of the eye with an ability to convert the light-ray activity ('image') into an electrical signal, which is transferred to the brain for decoding via the optic nerve. The central and most highly photo-sensitive region of the retina is known as the macula, which is crucial for visualising the detail of the image. It is comprised of the fovea: a pit-like structure near the macular region responsible for sharpest vision, and the optic nerve: a bundle of nerve fibres that form a nerve to carry the electric signal to the visual cortex of the human brain; e) the sclera: white structural tissue of the eye which consists of thick connective tissue supporting the wall of the eye, also protecting its internal structures, and f) the choroid: a layer of rich blood vessels packed between the retina and sclera which provides the nutrients to different cells in the retina.

1.2 Retinal architecture

The visual acuity of the human eye is highly dependent on the proper refraction of light travelling through the anterior structures before striking the retina [4]. The retina consists of a complex structure (Figure 1.2) of different neuronal cells, arranged between two layers of the retina, the anterior nerve-fibre layer, and the posterior retinal pigment epithelium (RPE). There are five major neuronal classes present in the retina: photoreceptors (rods and cone cells), bipolar cells, ganglion cells, horizontal cells, and amacrine cells, which pass on the electrical impulse to the brain through synaptic exchanges between them [5]. Light travels through the non-photosensitive layers of the retina before reaching the photoreceptors. The photoreceptor cells known as rods and cones are the only cells that contain photosensitive pigments called opsins (rhodopsin in rods and photopsins in cones), in the form of membranous disks in their outer segments, which transduce photons into an electrical impulse.

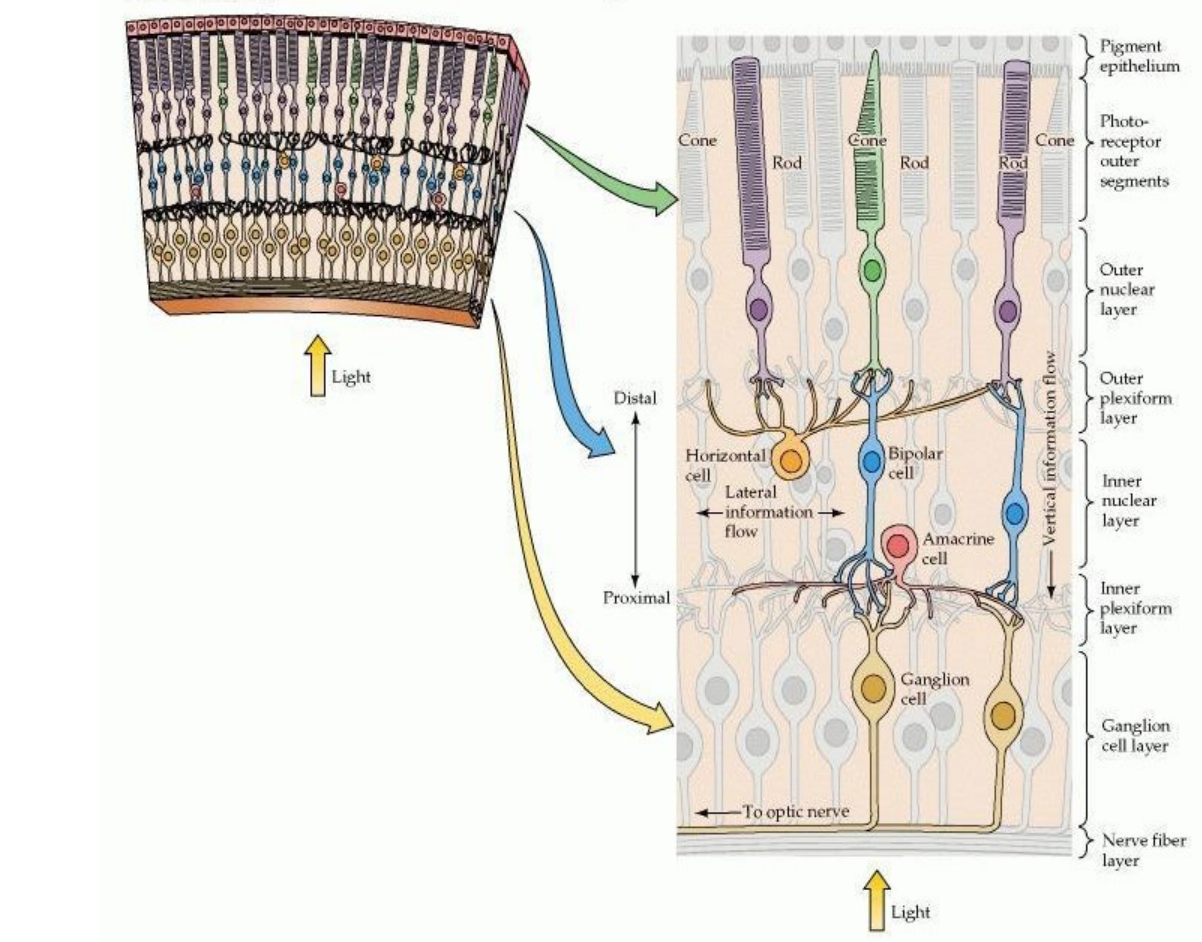

Figure 1.2: Cross-section of the retina showing the arrangement of neurons between the anterior and posterior retinal layers along with the direction of light source. Source: [5]

From the posterior side of the eye, the retina consists of layered cells as follows: the outermost layer is the retinal pigment epithelium (RPE) support tissue (detailed below), then the other layers existing sequentially are the photoreceptor outer segment layer, the outer nuclear layer, the outer plexiform layer, the inner nuclear layer, the inner plexiform layer, the ganglion cell layer and the nerve fiber layer. The inner nuclear, outer nuclear and ganglion cell layers contain the cell bodies of the retinal neurons while the synaptic exchanges take place in the inner plexiform and outer plexiform layers. The outer nuclear layer contains the cell bodies of the photosensitive cells while the inner nuclear layer contains those of the bipolar, horizontal and amacrine cells, known as the interneurons. Müller glia cells are the only non-neuronal cells present in the inner nuclear layer that assist in maintaining the

structural integrity of the retinal tissue [1]. The horizontal and amacrine cells help in the lateral flow of information in the retina while the bipolar cells exchange the signal with the retinal ganglion cells, whose cell bodies form the ganglion cell layer. The anterior-most layer of the retina is the nerve fiber layer, made of the axons of retinal ganglion cells. These nerve fibres converge to form the optic nerve that ultimately transfers the photo-stimulation signal from the retina to the central nervous system.

1.3 Phototransduction and visual cycle

The absorption of light by the photosensitive pigments in the neural retinal cells, rod and cone photoreceptors, termed 'phototransduction', initiates visual signaling [1], [7]. Rod cells are highly sensitive in dim light and saturate in brighter light. Cone cells are responsible for colour vision and high acuity with lower sensitivity to and thus functionality in bright light. The photosensitive pigment (opsin) is hyperpolarized by light absorption and initiates a cascade of events that culminates in neurotransmitter release ('the visual cycle', described below). Thus processed, the photonic signal is transduced as an electric potential through synapses in the outer plexiform layer with bipolar and horizontal cells. These secondary neurons in the retina are responsible for the later flow of information and a wide range of contrast over different photo-radiations. The photoreceptor cells receive feedback from the horizontal cells (provide feedback and feedforward signals to photoreceptors and bipolar cells by receiving glutamatergic inputs) causing either hyperpolarization, during exposure to light, or depolarization, in darkness, of the pigment membrane. The different sub-classes of amacrine cells, which are functionally diverse, act as information bridges in the inner plexiform layer, receiving signals from bipolar cells and transmitting information to the axons of retinal ganglion cells as well as contributing to visual function by the extension of visual signals laterally. Ganglion cells that form the retinal ganglion layer receive inputs from secondary neurons and respond to various stimuli such as colour, contrast and darkness. They actively propagate the action potential to the brain through the nerve fibres formed by the myelinated ganglion cell axons, which combinedly form the optic nerve [5], [6].

The visual cycle is a means of recycling the photosensitive chemical 11-cis retinal that mediates the first step in phototransduction. In photoreceptors, 11-cis retinal (a photosensitive derivative of vitamin A) is covalently linked to opsin (a G-protein coupled receptor) which then becomes the photosensitive pigment [1], [7]. The photoisomerization of 11-cis retinal to all-trans retinal causes a conformational change in opsin leading to membrane depolarization and generation of an electrical impulse by the photoreceptors. For the proper functioning and survival of photoreceptors, all-trans retinal must be converted back into 11-cis retinal with the help of a series of enzymatic reactions involving the retinal pigment epithelium [6]. The spent all-trans retinal is 'bleached' and travels to the outer segment of photoreceptors [8], where it is transferred to the RPE, after reduction to all-trans retinol. In the RPE, all-trans retinol is esterified by lecithin retinol acyltransferase (LRAT), further hydrolysed to 11-cis retinol and oxidized to 11-cis retinal [9], [10]. The restored 11-cis retinal is transported to the outer segment through the RPE and sub-retinal space and made available to the photoreceptors for regeneration of photosensitive pigment, thus completing the visual cycle.

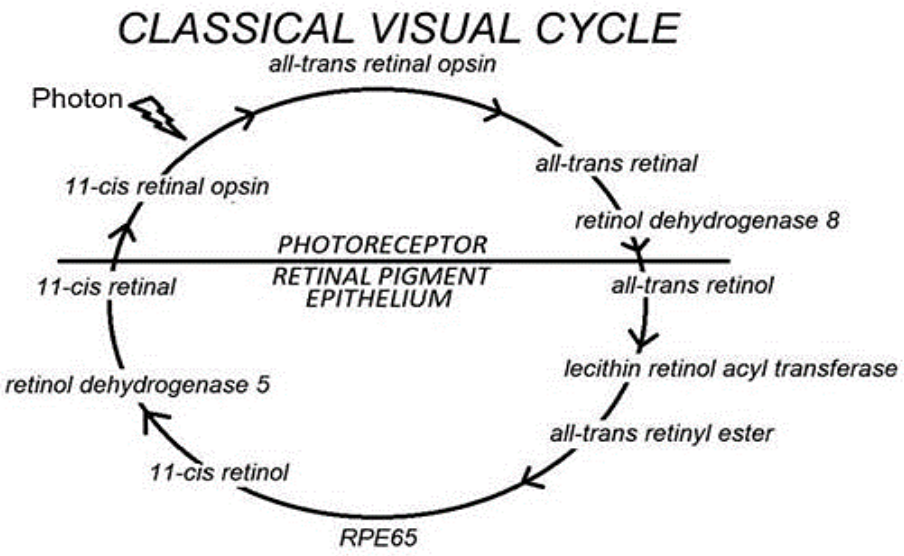

Figure 1.3: Illustration of the classical visual cycle. Source: [11]

1.3.1 Role of the Retinal Pigment Epithelium (RPE)

The spatial arrangement of retinal layers appears counterintuitive, as the light travels through all non-photosensitive layers and retinal vasculature before the photons are transduced. The structural organisation of the retina is explained by the interdependent relationship that exists between the outer segments of photoreceptors and retinal pigment epithelium (RPE, a dark pigmented layer that absorbs the transmitted light and reduces optical interference) for the exchange of energy substrates and assists in the viability of Retina [4]–[6]. The relationship between photoreceptors and the RPE is critical for numerous metabolic functions during the visual cycle and is essential for the normal function and survival of photoreceptors, although not an integral part of the neural retina. In addition to mediating the visual cycle (Figure 1.3), the RPE renders an essential role in removing the expended receptor disks by phagocytosis, at a rate of ~ 10% of the disks per day, as well as promoting the disk turnover by their renewal at the base of the outer segments [6] and maintaining the blood-retina barrier, composed of retinal capillary endothelial cells (inner layer) and retinal pigment epithelial cells (outer layer), that constantly regulates the nutrient flux [12].

Literature Review

Genetic variants contribute to people's risk of developing many diseases. The reference human genome sequence, provided by the Human Genome Project (HGP), has proven to be a powerful resource for expanding the study of human genetics. This resource increases the power of gene discovery and diagnosis of diseases (common and rare) through genomics - the study of all genes and their interactions in a person/population [13]–[16]. Studying the heritable variation in a trait, due to contributions from multiple genetic variants, at a population scale, contributes to our understanding of the functional aspects of the human genome, which is crucial to elucidate the mechanisms of expression of a trait/disease. Genetic polymorphisms such as single nucleotide polymorphisms (SNPs), insertions and deletions (indels) and larger structural variants (duplications/inversions/expansions), together termed ‘variants’, play a pivotal role in the development of disease and can thus reveal underlying

biological mechanisms [17], [18]. The primary objective of genomic studies is to locate the DNA variants that influence a particular trait (including disease traits) that are significantly over or underrepresented among the affected individuals relative to the general population. Measuring human DNA variation and gene expression are essential for carrying out genomic research. Since the completion of the Human Genome Project, an expanding wealth of catalogued human genetic variation is available in repositories of single nucleotide variations (SNVs) including the Thousand Genomes Project (TGP), dbSNP and gnomAD. dbSNP and TGP helped in developing SNP arrays. GnomAD [19], a comprehensive database of human variation aggregated from numerous studies, aids in improving statistical methods to characterize heritable genes specific to a trait/phenotype and generate a comprehensive understanding of disease aetiologies. ClinVar is another important resource for all types of human variations (germline and somatic) providing functional evidence along with their genomic location and clinical consequence. [17], [20], [21]–[23].

1.4 Technologies for exploratory genetics

DNA genotyping

Profiling common genetic variants requires DNA genotyping. SNPs are the predominant DNA polymorphisms, which are single base polymorphisms (variations) that occur approximately every 1 in 1000 bases compared to a reference genome, either by transitions (C/T or G/A) or transversions (C/G, C/A, or T/A, T/G), at a single nucleotide position [17], [24]. SNPs are catalogued in databases such as GnomAD and their frequency in populations varies considerably, with many rare variants (<1% minor allele frequency (MAF)) than common variants of the 241 million small nucleotide variations [19]. Genotyping is the process of determining differences in the genetic make-up (generally restricted to SNPs) of an individual on a high-throughput platform. This can be performed either in a targeted manner or using a genome-wide approach, with the help of known signature SNPs (tagging SNPs) considering linkage disequilibrium (LD, non-random association of alleles/SNPs [25]) structure into account. The index population of commercially available SNP arrays is generally based on TGP [26].

Genotype Imputation

The availability of high-quality reference genomes for thousands of individuals of different ancestries means it is feasible to genotype only a small fraction of genetic variants of a single sample using a SNP array. As only a subset of the genetic variants are genotyped, Genotype Imputation employs statistical methods with a probabilistic framework, by exploiting the LD between SNPs, to ascertain the unobserved genotypes by comparing them to high-density genotype reference panels [27], [28]. The major advantages of imputation are the increased power of subsequent genome-wide association studies (GWAS, detailed below) [22], [29], [30], decreasing the number of SNPs that must be directly assayed/genotyped, and providing robust data for fine-mapping and meta-analysis. Hence, this methodology has become ubiquitous in modern computational genetics/genome-wide studies, since the essential reference panels could be generated with the public availability of genetic variation information from various large-scale studies (due to ongoing improvement of SNP arrays as well as a reduction in sequencing costs) such as the International HapMap Project [31]–[34], the 1000 Genome Project [17], the UK10K project [35], Haplotype Reference Consortium (HRC)[36] and the Trans-Omics for Precision Medicine (ToPMed) program. Various compute-intensive imputation tools are available, both standalone and web-based, for performing genotype imputation with high quality. IMPUTE2 [37], Beagle [38] and Minimac4 [39] (web-service: Michigan Imputation Server) are a few amongst them.

RNA sequencing

Messenger RNA (mRNA) transcripts are produced from genes via DNA through the process of transcription. To quantify all transcripts, RNA is extracted, fragmented, copied and sequenced in a high-throughput manner termed RNA-seq [40] typically using short-read sequencing platforms, with Illumina being the predominant platform currently in use. The gene expression is quantified by mapping reads to the human genome reference using tools such as STAR [41] and then counting the number of reads (read count) that align to each gene using computational tools like HTSeq-count [42] and featureCounts [43]. RNA-seq data can also facilitate the discovery of novel transcripts [44], performing differential expression

analysis studies [45], and identification of alternatively spliced genes [46]. Combined with genetic variants, RNA-seq can further allow detection of allele-specific expression [47] and eQTLs analysis (detailed below).

DNA-seq is also possible with both whole genome and whole exome sequencing, with both technologies delivering a substantial increase in genomic diagnoses, in particular for Mendelian genetic disorders [48], [49]. Neither of these methodologies are used in this thesis and are mentioned for completeness.

1.5 Statistical genetics and association studies

The wealth of genetic data from RNA-seq, DNA-seq and DNA genotyping are analysed with a variety of statistical methods, which have greatly increased our ability to interpret both simple Mendelian and complex, multigenic phenotypes. Common complex diseases arise due to the combined effect of multiple genetic variants, each with modest disease contributions. Individuals, including both cases and controls, have a random selection of these variants which yield an overall contribution to disease risk. These diseases occur with lower penetrance and are also influenced by environmental factors such as diet, gender, age etc. In contrast, simple rare genetic diseases are usually the result of single genetic variations, showing high penetrance. Penetrance is defined as the proportion of individuals containing the variation in a gene and expressing the gene-related trait [50]. Variants associated with common complex and simple rare genetic diseases, hereafter will be referred to as ‘common’, and ‘rare’ genetic diseases respectively. The penetrance of these disorders also leads to a general trend of more permissive frequencies for common complex disease risk variants in comparison to the rare monogenic diseases, which are almost always extremely rare alleles. This is depicted in a standard representation comparing allele frequencies of risk variants versus one of penetrance, genetic risk or effect size (Figure 1.4).

1.5.1 Linkage analysis

The genetic location of disease-causing variants can be evaluated using family-based linkage-mapping, or population-based case-control studies. Rare Mendelian diseases are

caused by genetic variations with greater penetrance in a population and are discovered by linkage analysis. Linkage analysis determines if a variant is physically linked (i.e., proximal on the same chromosome) with the disease gene within a family, by testing for co-segregation of known disease risk genetic variants with the help of genotype data from closely related family members [51]. Linkage analyses are aimed at finding deleterious rare variants, but lack the power to identify disease susceptibility loci with more subtle contributions to common genetic diseases [52], [53].

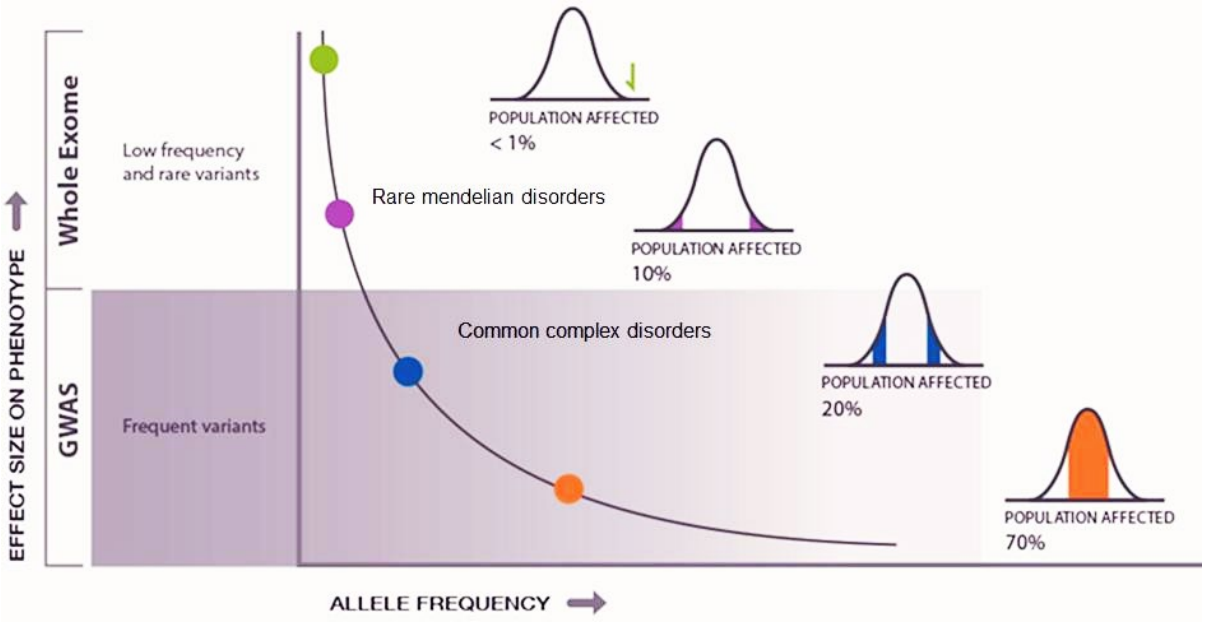

Figure 1.4: Contrasting the frequency and effect size of disease influencing variants. GWAS studies are the most powerful low effect size, high-frequency variants in contrast to linkage analysis which is a powerful approach to identify disease-causing variants with a large effect size that are generally rare (have low allele frequency). Source: https://www.gbhealthwatch.com

1.5.2 Genome wide association studies

Population-based case-control studies, wherein the frequency of genetic variants in a group of individuals with the disease (cases) is compared to that in individuals without the disease (controls), known as Genome-Wide Association Studies (GWAS), are used to identify common variants that underlie complex traits. The essential components of a GWAS

study design (Figure 1.5) are: a large sample of cases and controls, genotyped with a high throughput assay, such as a SNP chip, demographic data and information on any potential technical confounding factors. For studying continuous traits, linear regression-based methods are widely employed to perform association tests to identify genetic associations without bias [54]–[56]. For binary traits, logistic regression methods are used instead. SNPs are the most common genetic markers that are currently obtained easily on a genome-wide scale. The term "phenotype" (a.k.a trait; in many cases, disease) denotes any measurable characteristic determined by the underlying genotype. GWAS on gross, often physical phenotypes, measured as a quantitative trait, are technically similar to eQTL studies (expanded upon below), where intermediate phenotypes such as gene expression are the trait of interest. The analysis of GWAS datasets can be performed using a program called PLINK [57], a versatile toolkit with modules for data organization, formatting, quality control, and association testing. However, a series of more sophisticated methods like SAIGE (Scalable and Accurate Implementation of GEneralized mixed model) [58] have been developed in recent years that extend these methods, to, for example, include related individuals.

The statistical power of GWAS to detect SNP-trait associations depends on multiple factors such as the experimental sample size, the frequency of variants, the effect sizes of the genetic variants that are segregating in the population, and the LD that exists among the genotyped variants. The resulting P-values from an association study are generally visualised on a genome-wide level using a Manhattan plot. Those SNPs with the lowest P values are positioned at the top of the plot, representing the highest disease association. The plot has an appearance of the Manhattan skyline and is thus called a Manhattan plot. The x-axis of this plot represents the regions on the genome (chromosomes) and the y-axis represents the negative log-transformed P-value (Figure 1.5). Because each SNP is independently tested for disease association, and multiple testing correction is difficult in the presence of LD, as a rule of thumb P-values < 5E-8 are termed 'genome-wide (GW) significant' with P-values < 5E-6 being termed 'suggestive-significant'. Associations at the GW-significance threshold are highly replicable [22].

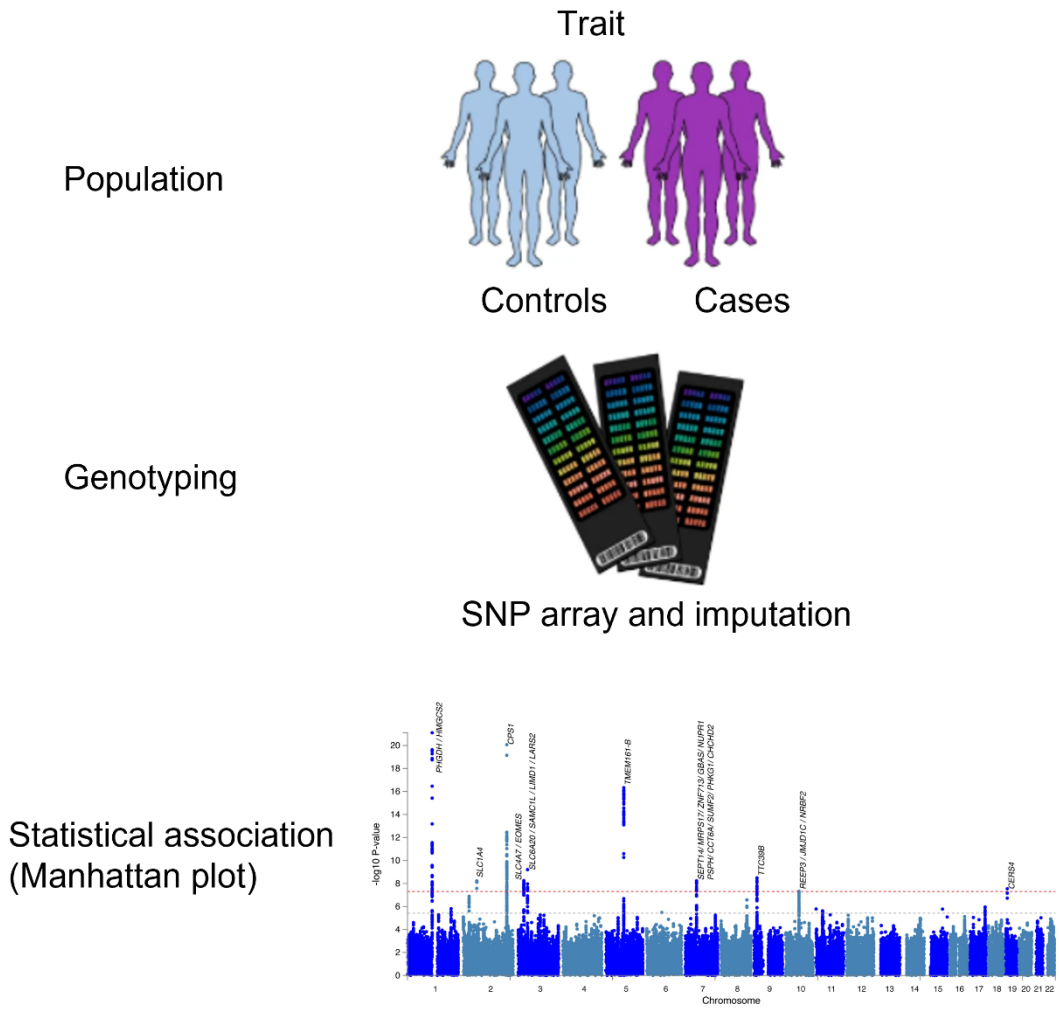

Figure 1.5: Schematic showing a GWAS study design with Manhattan plot of the P-values for each locus generated from the GWAS.

1.5.3 Post-GWAS methods aid biological interpretation of results

The causal variant or variants (SNPs that collectively contribute to the manifestation of a trait/disease) are frequently not identified by GWAS due to LD and genotyping coverage/imputation accuracy [59]. Functional characterization of such candidate variants is therefore imperative to identify the disease-specific biochemical mechanisms and causal variants, especially when association regions lie outside protein-coding regions of the genome, which is frequently the case [60]. DNA features such as promoter regions, and

transcribed variants in introns, and long and short ncRNAs are known to affect protein-coding gene expression [61]. Performing eQTL analysis (expanded below), is a powerful approach to identify biochemical mechanisms stemming from the association between genetic variants and gene expression.

In recent times, the Functional Mapping and Annotation of Genome-Wide Association Studies server (FUMA-GWAS), an easy-to-use web-based platform, has emerged as a prioritizing tool by integrating 14 state-of-the-art biological data repositories to reveal the potential causal variants providing a way to interpret and functionally annotate the GWAS summary statistics. This tool can aid the identification of biological implications of the associations through the two core functions SNP2GENE and GENE2FUNC which annotate the summary statistics to prioritize causal genetic variants and provide tissue-specific gene expression patterns respectively [62]. FUMA includes other functional analysis tools such as chromatin interaction mapping, used to map SNPs to genes based on a significant chromatin interaction between the disease-associated loci and nearby or distant genes for a selected tissue/cell type; integration of MAGMA [63] for gene analysis and independent gene-set analysis for convergence of prioritized genes in biological function/tissue as well as in Genotype-Tissue Expression (GTEx) database for tissue-specific prioritization of SNPs. Crucially, however, FUMA currently does not utilise retinal eQTLs.

1.6 eQTL mapping

Quantifying genetic effects on intermediate phenotypes such as gene and protein expression are essential for a mechanistic understanding of GWAS results. Despite the publication of thousands of GWAS results that indicate links between genomic loci/variants and complex traits, the mechanistic links between these variants and phenotypes are often opaque. In fact, the majority, ~ 93% of significant variants are located in non-coding regions of the genome [64], [65]. The method known as ‘quantitative trait locus’ (QTL) mapping can bridge the effects of genetic variation on phenotypic variation. ‘eQTL’ denotes two entities: ‘eSNPs’ and ‘eGenes’. ‘eSNPs’ affect the expression of a proximal ‘eGene’. The phenotypic association of an eSNP allows us to estimate the effect size (magnitude of its contribution) for an eGene [66]. Any GWAS hit could be a potential eSNP or located in close proximity

(or in LD) to a contributing eSNP. This marker/locus may also be called an eQTL hotspot if it affects the expression of multiple genes downstream [67]. As for GWAS, non-genetic/environmental factors like gender, weight, blood pressure, are included in the analysis to mitigate confounding of the eSNP-eGene relationship. eQTL mapping studies prioritise genetic variants (eSNPs) identified within a GWAS locus, which can function in a tissue-specific manner. These eSNPs can affect transcriptional enhancers (sites of transcription factor binding), among other physical genetic interactions that affect expression in a tissue [68], [69]. Initial eQTL studies performed using Lymphoblastoid cell lines (LCLs) across diverse populations from the HapMap3 project, found that eQTLs are largely (at least half identified) shared amongst human sub-populations/ethnicities [70]. In a relevant example to the retina, a recent study on conserved non-exonic elements identified that miR-9 (a microRNA) depletion affects retinal vasculature formation and found that transcriptional enhancers can be disrupted by conserved non-coding SNPs [60]. Also, it is observed that the eQTL effect size, as well as the direction of the effect, are conserved across populations and that variants close to the transcription start site (TSS) have a stronger effect on the neighbouring genes [70]. The premise of eQTL studies generally is, by combining and exploiting the stable DNA data and informative RNA data, they can provide more evidence to understand complex diseases.

1.6.1 Types of eQTLs

eQTLs are divided into *cis*- and *trans*- subtypes. *cis*-eQTLs (Figure 1.6), are located on the same chromosome, close to the affected gene (showing local effects, e.g. conventionally defined as being located within 1 Mb distance) whereas *trans*-eQTLs are located distal to the affected gene (showing distal effects, e.g. >1 Mb distance, or on another chromosome) [71]. *cis*-eQTLs are close to the TSS and generally show relatively large effect sizes (e.g., as seen in Figure 1.6, the expression level of *cis*-eQTL almost doubles with changes in genotype from AA to BB).

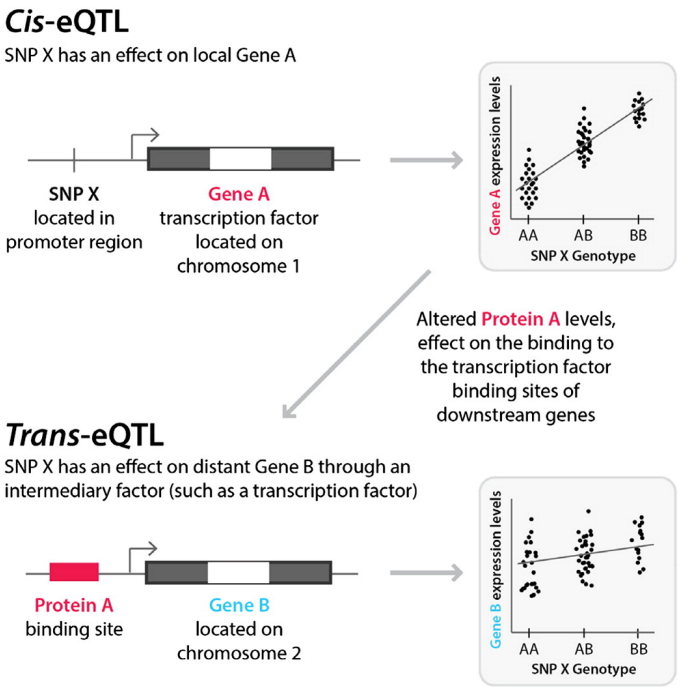

Figure 1.6: Types of eQTLs showing the allele-specific expression in terms of distance from the affected gene. Source: [71]

eQTL analysis presents a statistical testing challenge due to the large number of statistical tests that need to be performed. With a sample size of N individuals with M markers, complete testing of all eQTLs would result in N times M statistical tests (compare for GWAS where there are typically only M tests, and even there the statistical thresholds for significance due to multiple testing correction requirements are very stringent).

In *cis*-eQTL analysis the statistical correction required for multiple comparisons is less stringent than for *trans*-eQTLs, as only markers near the affected gene are considered, leading to far fewer statistical tests needing to be performed. Whereas in the case of *trans*-eQTLs greater numbers of SNPs covering the entire genome are considered, which results in excessive false positives due to a severe multiple testing problem [72]. Finding *trans*-eQTLs is much more challenging and requires a larger sample size to achieve similar power or effect sizes compared to *cis*-eQTLs [73]. Accordingly, *trans*-eQTL analysis methods require further development to produce reliable results, in particular through prioritisation of which trans-eQTLs to test. Although less is known about the *trans*-eQTLs effect on distant genes,

a study based on statistical mediation analysis (using Sobel tests of mediation, Sobel P < 10e-5 [74]) have provided evidence (using a cohort of 1800 South Asians) that ~ 35% of the *trans*-associations are significantly mediated by a local (*cis*) transcript (*cis*-eGene expression) [73], [74].

1.6.2 Computational tools for eQTL analysis

eQTL mapping is performed by integrating and statistically testing the genotype data and the gene expression abundance for a cohort using computational algorithms. These are available either as a standalone and/or collections of tools available via web servers for performing high throughput eQTL analysis using omics data. Popular software options are described in Table 1.1. The tools are either based on linear regression (often a simple model with an additive gene effect to increase power [75]) or non-linear models [76]. The most popular computational tools are based on linear models such as, QTLtools [77], MatrixeQTL [78] and PLINK [57], with versatile applications that fit a linear equation between the gene expression data (the independent or outcome variable) and genotype (dependent or regressor variable), denoted 0:homozygous reference 1:heterozygous or 2:homozygous alternative. Tools such as GEMMA [79], based on the Linear Mixed Models (LMM), are optimized to account for polygenetic effects and to minimize false positives. Further, tools such as MT-HESS [80] and HT-eQTL [81], and iBMQ [82] implement Bayesian mixed models. Other tools based on non-linear models include Merlin [83], eQTL-LMT [84] and R/qtl [85]. However, the visualisation functionality is lacking from these tools and requires either custom visualisations (e.g., Allele-specific boxplots) or use additional tools like LocusZoom [86] to visualise eQTLs in a locus. Despite the availability of many eQTL tools, given the complexity of the omics data with other confounding factors such as population structure and heterogeneity, further memory-efficient approaches reducing the computational time to identify the significant true positive eQTLs are still necessary.

Table 1.1: A collection of popular computational eQTL analysis tools.

Software	Platform	URL	Model	Strengths	Limitations
eMap [87]	R	bios.unc.edu/~weisun/software.htm	Linear regression	• Includes a visualization and fine mapping module • Useful to identify eQTL hotspots.	• No support for Windows • More suitable for small-scale analysis • Does not handle covariates
QTLtools [77]	C++ Command-line	https://qtltools.github.io/qtltools/	Linear regression	• A complete tool kit for both cis and trans eQTL analysis with ultra-fast methodology. • It has all the modules from pre-processing to post eQTL fine-mapping analysis with the ability to handle all common genomics file types. Improvement of FastQTL [88].	-
Matrix eQTL [78]	R, MATLAB	bios.unc.edu/research/genomic_software/Matrix_eQTL/	Linear regression	• Designed for both cis and trans eQTL analysis of large datasets • Fast without loss of precision • Supports multiple statistical models	-
PLINK [57]	C/C++, Command line	www.cog-genomics.org/plink2	Linear regression	• A versatile tool supporting all the pre-processing steps such as data manipulation, filtering etc. • Suitable for both small- and large-scale analysis	• Not optimised for eQTL analysis • very time cumbersome when covariates are included in the model
BootstrapQTL [89]	R	https://github.com/InouyeLab/bootstrapQTL	Linear Mixed Model	• Implements bootstrap procedure to correct the overestimation of effect sizes	• Depends on MatrixEQTL internally for eQTL mapping

GEMMA [79]	C++	https://github.com/genetics-statistics/GEMMA	Linear Mixed Model	• Includes multiple linear mixed models for fast association tests in GWAS with estimates for variance components	• Mostly GWAS concentrated method and not specific for eQTL analysis
MT-HESS [80] HT-eQTL [81]	C /MATLAB	http://www.mrc-bsu.cam.ac.uk/software	Hierarchical Bayesian	• Good for eQTL hotspot discovery and first tool to investigate the systemic role of a genetic marker • Scalable multi-tissue method for eQTL analysis • Increases power of eQTL detection in a single tissue by borrowing strength across tissues.	• May not suit a single tissue eQTL or non-differential based analysis
iBMQ [82]	R	http://bioconductor.org/packages/release/bioc/html/iBMQ.html	Hierarchical Bayesian	• Improves detection of large scale trans-eQTL hotspots • Computationally efficient for large datasets	-
MERLIN [83]	C/C++	https://bioinformaticshome.com/tools/gwas/descriptions/Merlin.html	Non-linear	• Tool for fast pedigree analysis but with an option for analysing unrelated individuals • Useful for simulation studies	• reduced performance in large-scale analysis • More suited for pedigree analysis • Not maintained anymore
eQTL-LMT [84]	R Python/Shell	https://github.com/beretta/eQTL-LMT	Logistic Model Trees Machine learning	• Supervised machine learning approach for integrating the eQTL predictions from different tools to improve eQTL mapping	• requires the usage of other eQTL tools beforehand and compute-intensive
R/qtl [90]	R	rqtl.org/	Hidden Markov Model	• eQTL analysis package that can deal with missing genotype data • Contains built-in functions for quality control checks as well as to estimate genetic maps.	• Exclusive to R

1.6.3 eQTL repositories

The Genotype-Tissue Expression (GTEx, URL: https://www.gtexportal.org/home/) project is a comprehensive public resource that provides data on tissue-specific gene-expression and eQTLs [68] including visualization browsers for gene-expression data. Currently, genotype data is available for 54 tissues (data generated from 838 donors with 15,201 samples), of which 49 have cis- and trans-eQTL data available (with at least 70 samples with both whole genome sequencing and tissue-specific bulk RNA-seq data), providing a huge atlas of the gene regulatory landscape including splicing QTLs (sQTLs, the variants associated with different RNA isoforms expression) for the majority of tissues [91]. GTEx contains 4,278,636 genetic variants associated with gene expression. GTEx has emerged as a primary resource to query for the functional interpretation of GWAS associations providing remarkable detail of gene expression regulation with at least one cis-eQTL for 94.7% of all protein-coding genes [91], [92]. Further, it includes information about mechanistic insights of complex traits accommodating for the influences of gender, ancestry and cell-type composition. Another public resource that has become recently available is the eQTL Catalogue [93] of the European Bioinformatics Institute, which hosts eQTL data as well as splicing QTLs uniformly recomputed from 21 independent eQTL studies. Currently, however, the GTEx project does not contain the eQTL data for the retina among the 54 tissues. Of the two recently published retinal eQTL studies, at present the data from one of these studies that mapped eQTLs in retinal tissue for both controls as well as in AMD patients is available as EyeGEx, an external data set linked to by GTEx [94].

1.6.4 Statistical power for eQTL mapping

In the post-GWAS era, eQTLs are generally studied through a genome scan approach to identify the eSNP:eGene relationships for a tissue or disease phenotype and locate a trait locus containing the major eQTL effect [95]. Statistically, eQTLs are mapped by associating the mRNA abundance for a gene with genetic variants to pinpoint a locus regulating or explaining the expression of a transcript at a population scale. Determining the statistical power to reject a null hypothesis correctly, and to detect a biologically meaningful eQTL signal from the study dataset is a crucial step during the design phase of genome-wide eQTL

studies. For eQTL studies, the null hypothesis is the absence of a genetic variant's effect on gene expression and therefore power is defined as the probability of detecting a true positive eQTL or rejecting the null hypothesis when true association exists [96], [97]. As a rule of thumb, a study design is considered good if it reaches a power of at least 80%. Researchers cannot control or even predict the actual genetic architecture underlying a phenotype, such as expression. However, the power of the study can be enhanced by determining the range of effect sizes detectable at a specific p-value threshold, typically set at 0.05, after controlling for multiple testing. Hence power studies can be performed under scenarios such as (i) determining the effect size achievable, given a set of eQTL tests and a particular sample size, or, (ii) identifying the sample size necessary to determine transQTL effects with at least a minimum threshold effect size, as well as other scenarios [97]. A recent study provided insights about the power of detecting eQTLs through realistic simulations of eQTL data (six sample sizes ranging from 100 to 5000) for various important statistical power determining factors for eQTLs such as the Minor Allele Frequency (MAF), the effect-size (beta), sample size and multiple testing correction procedures [89]. The authors concluded that eQTLs identified from lower sample sizes with small minor allele frequencies are potentially false positives. A statistical power of 80% with an effect size of 0.25 s.d. per allele could be attained at a sample size >1000 and MAF >25% and infer, that studies with 100 or fewer samples are underpowered to detect eQTLs even at MAF >25% unless they have large effect sizes. Therefore, sample size plays a crucial role in detecting true positive eQTL associations. Another recent study has developed a state-of-the-art power calculator for bulk tissue and single-cell eQTL analysis, known as powerEQTL [98]. This R package and an interactive webtool implements two different statistical models such as one-way unbalanced ANOVA and simple linear regression. This versatile application can aid in estimating the power of a study, given a sample size, effect size and minimum MAF or any of the four parameters if three of them are specified by the user.

1.6.5 Limitations of eQTL mapping: the need for post-eQTL methods for fine mapping of variants

Although eQTL mapping expands our understanding of the architecture of gene regulation, the non-independence of SNP segregation (LD structure) and the high

dimensionality of gene expression data complicates our ability to derive clear insights about the molecular mechanism of a quantitative trait. Additional functional studies are therefore indispensable to fine map the eQTL signal to enhance our understanding of the eSNP:eGene relationship.

Colocalisation analysis

Colocalisation testing is an integrative approach to further understand GWAS associations by comparing GWAS signals from multiple traits, which may or may not be derived from the same cohort, including expression data. The basic methodology of this approach is to perform a correlation analysis on the summary statistics of multiple GWAS studies with independent eQTL studies, to test whether the association signals are consistent in their effect size relationship. If this were the case this would indicate that the same haplotype (LD signal) underpins the two traits suggesting that the same driver variant is associated with both traits, even if that driver variant remains unknown. Such integrative approaches using eQTL summary statistics help reveal the functional effects of GW-significant variants and aid in understanding the SNP:gene causal relationship pertaining to a GWAS locus [99], [100]. There are multiple colocalisation tools, for instance, *coloc [99]* and eCAVIAR [101], developed by employing either a probabilistic or Bayesian statistical framework to interrogate the probability of colocalisation. *coloc* is based on an approximate Bayes factor (Bayesian method) approach to test if the same underlying variant is responsible for the signal by testing GWAS-GWAS and GWAS-eQTL summary statistics. *coloc* estimates the posterior probabilities to identify if the associated signal is distinct or shared by the studies under consideration by assuming only a single causal variant at a given locus. Notably, the power of *coloc* results may be reduced when multiple causal variants exist [100]. eCAVIAR is another methodology based on a probabilistic framework. It estimates the colocalisation posterior probability at a given locus, accounting for multiple causal variants. The most recent GWAS for various tissues including retinal disorders have utilised this approach to identify shared associations between traits. A recent eQTL meta-analysis on brain regions, a tissue closely related to the retina, included colocalisation analysis using *coloc* by integrating variants from schizophrenia GWAS summary statistics have identified seventeen genes out of 128 independent GWAS signals for schizophrenia, that achieved a

posterior probability > 0.7, providing useful information to further understand gene regulation in the brain [102].

Transcriptome-wide association studies

Identification of eQTLs enhances our ability to establish robust gene-trait associations via approaches like transcriptome-wide association studies (TWAS) [103]. TWAS involves using multiple eSNPs calculated in independent studies in disease-relevant tissues, to construct models that predict the expression of each gene based on the disease genotype (summary statistics). Pre-computed SNP:gene weights for non-retinal tissues are available through the FUSION software [104] (Figure 1.7). Genetically predicting gene expression for cases and controls for a trait of interest can predict genes likely to show trait-related differential expression. Such analysis enhances our power to identify disease-risk genes, particularly when it is not easily feasible to obtain anatomical specimens at a population scale (e.g., for the retina) and also to mitigate the actual costs involved in direct RNA sequencing.

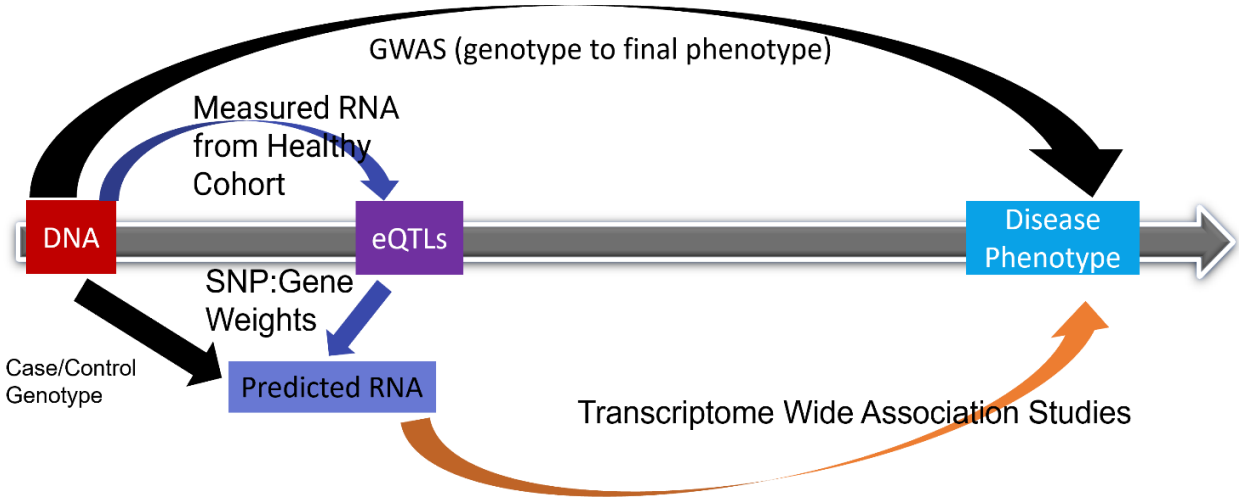

Figure 1.7: A schematic representation of TWAS. It denotes that mapping eQTLs will enhance the power to predict the gene expression with genotype alone and applying TWAS can lead us to identify gene-phenotype associations.

1.7 Genetics and population studies of complex retinal diseases

Many genetic retinal disorders are debilitating conditions resulting from the degeneration of retinal cells, which can progress to blindness. The symptoms of different retinal disorders include blurred and distorted vision, loss of side and/or central vision as the photoreceptors and/or surrounding cell types malfunction. These are either genetically based systemic diseases that affect the retina; or rarer non-systemic genetic diseases affecting the macula, which mediates advanced visual function. Complex disease aetiology can also be influenced by environmental and lifestyle risk factors such as smoking, diet and alcohol intake among others [105]. Age-related macular degeneration (AMD), retinitis pigmentosa, diabetic retinopathy, retinal detachment, and macular telangiectasia (MacTel) are complex retinal disorders that together exert significant global health burden. The cost of vision loss due to all causes globally (as per [106]) is estimated at US$3 trillion, while that of AMD alone is US$343 billion. Of primary interest in this thesis is the disease aetiology of two retinal disorders: AMD, and Mactel. AMD (prevalence ~ 8-9%) is vastly more prevalent than MacTel (~ 0.1%) but both diseases feature complex genetic architectures which may also interact with environmental risk factors. The following sections will describe common statistical approaches for understanding complex diseases, and the results achieved for AMD and MacTel.

1.7.1 Age-related Macular Degeneration and GWAS studies

AMD is the leading cause of irreversible adult blindness globally [107]. AMD is an acquired visual impairment that generally occurs after ~ 60 years of age with a prevalence varying by ethnicity, with the greatest prevalence of 12.3–30% observed in individuals with European ethnicity, and ~7.5% among African and Asian ancestry individuals [108], [109]. The symptoms of AMD include blurred, distorted and fuzzy vision, leading to complete blindness in some cases [108]. In AMD, the central most part of the retina responsible for central vision, the macula, degenerates progressively. Drusen (yellowish lipid deposits in the retina) are considered a hallmark of AMD (Figure 1.8b). Reticular pseudodrusen are also emerging as hallmarks of severe AMD [110]. The two phenotypes of AMD are atrophic (dry) and neovascular (wet) AMD, with dry AMD (85-90%) being more prevalent than wet AMD

(10-15%). Dry AMD is characterized by 'geographic atrophy' resulting from the death of photoreceptors and retinal pigment epithelium, while wet AMD is primarily due to the abnormal growth of new blood vessels under the macula. Both conditions cause severe loss of vision. Although the aetiology of AMD is incompletely understood, an abundance of genetic evidence from family, sibling and twin studies have established genetic predisposition and other non-genetic factors such as age, smoking, and pre-existing medical conditions as contributors to disease manifestation. AMD is now understood to be a complex genetic disorder [111], [112] its increasing prevalence [113] requires deeper disease understanding to promote new therapies.

AMD GWAS studies

The very first GWAS success in 2005 was the independent discovery from three cohorts of the association between complement factor H (*CFH*) *[114]* with AMD. In 2020, Han and colleagues published a GWAS that identified a total of 46 genetic risk loci [115], containing 69 independent AMD-associated variants. This study performed a meta-analysis, by integrating the summary statistics from previous studies, specifically: by the Age-related Macular Degeneration Genomics Consortium (IAMDGC) that compared 16,144 AMD cases with 17,832 controls; a GWAS in the Genetic Epidemiology Research on Aging (GERA) cohort with 4,017 cases and 14,984 controls; and three additional previous GWA studies which consisted of >17,100 cases and >60,000 controls of European and Asian ethnicity. This meta-analysis resulted in identifying 12 novel AMD associated loci. Variants in the complement factor H (*CFH*) gene (locus 1q32) have been associated with an increased risk for AMD progressing to bilateral geographic atrophy, while those in high-temperature requirement A serine peptidase 1, *HTRA1/ARMS2* contribute to bilateral choroidal neovascularization. Association of additional complement genes, *CFB*, *CFD*, *C2*, *C3* and *C5,* imply a significant role for the innate immune system in AMD pathogenesis [116], [117]. The Y402H amino acid substitution-coding variant in *CFH* reduces the binding ability of the *CFH* protein with other proteins such as heparin, leading to higher AMD risk [118], [119]. Variants in the age-related maculopathy susceptibility 2 protein (*ARMS2*, 10q26) and the high-temperature requirement A serine peptidase 1 (*HTRA1*, 10q26) genes, have been

mapped as the second major locus contributing to AMD [120], [121]. These two major loci explain over half of the heritability of AMD which is considered to be the most genetically well-defined of all complex disorders [122]. The heritability estimate (h^2) of AMD ranges from 0.44 to 0.62 differing with the disease severity [123]. The vascular endothelial growth factor A (VEGFA) causes angiogenesis, promoting fluid leakage from retinal blood vessels due to vascular permeability in wet AMD, an advanced stage [120]. Also, adjacent to these loci are genes with important biological functions, such as extracellular matrix genes (*COL4A3*, *MMP19* and *MMP9*), an ABC transporter linked to high-density lipoprotein (HDL) cholesterol (*ABCA1*) and a key activator in immune function (*PILRB*) [124]. Furthermore, mitochondrial dysfunction is another factor in AMD, causing an energy crisis for the retina and mtDNA polymorphism is one of the powerful predictors of AMD [120], [125]. Interpretation of GWA variants and attributing causality remains a major challenge as the individual effect sizes are small. To address this, a recently performed eQTL analysis with a cohort of 105 healthy and 348 AMD (total 453) retinas, which also established the Eye Genotype Expression (EyeGEx) database, provided important information for this tissue which was not previously part of GTEx [94]. cis-eQTLs mapped for retina comprised 14,565 eSNPs influencing the expression of 10,474 eGenes. Integrating retinal eQTLs with 24 AMD GWAS loci resulted in nine lead SNPs that attained statistical significance of the 19 retinal eSNP-eGene associations. Additionally, by employing transcriptome-wide association models, three genes: *RLBP1*, *PARP12*, and *HIC1*, were identified as the strongest novel candidate genes for AMD. Further independent studies are required to validate these most recent findings.

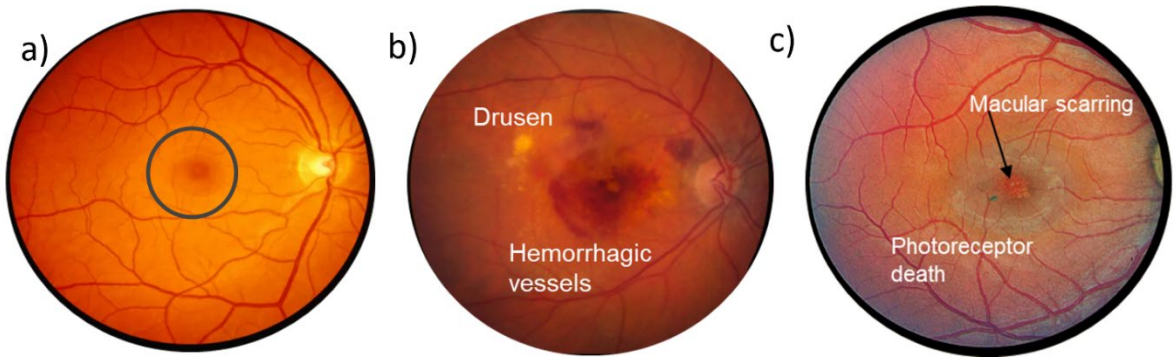

Figure 1.8: Retina fundus photographs of a) healthy retina, b) AMD affected retina and c) Macular Telangiectasia type II (MacTel) affected retina. The blue circle denotes the macula in the healthy retina. The AMD affected retina shows dark spots in the centre, likely haemorrhagic blood vessels and yellow drusen deposits damaging the macula. The MacTel affected retina shows a grey patch due to the photoreceptor cell death and macular scarring in the central region. Source:[126]

1.7.2 Macular Telangiectasia Type II and GWAS studies

MacTel is a rare neurovascular degenerative retinal disorder, with a prevalence of ~ 0.1%. The mean age at diagnosis is 60 years with symptoms such as blurred vision, slow dark adaptation and progressive vision loss (slow progress to more significant loss of central and/or paracentral vision) [127], [128]. It is a complex heritable disease with abnormalities developed in and around the fovea, which has the highest density of cone photoreceptor cells. The ophthalmological imaging characteristics of MacTel (Figure 1.8c) include macular scarring, decreased macular pigment, non-drusen crystalline deposits in the retina, retinal haemorrhages with widespread leakage of blood vessels and abnormal growth of vessels into the fovea which permeate the outer retina [129]–[131]. The cause of this bilateral retinal disorder is not completely known. Studies are ongoing to find effective therapies for MacTel patients and advances in our understanding of its genetic basis.

MacTel GWAS studies

The genetic architecture of MacTel was established by the first-ever MacTel GWAS with 476 patients and 1,733 controls (hereafter referenced as the 2017 GWAS), performed as a part of MacTel Consortium [127] which identified three GW-significant and two suggestive-significant loci. The primary locus enclosing the candidate gene transmembrane protein 161B (*TMEM161B*), is involved in retinal vascular calibre while the other two GW-significant loci were associated with the genes *PHGDH*, *PSPH*, and *CPS1*, which are central to glycine-serine metabolism. The latest subsequent MacTel GWAS study with a cohort of 1067 MacTel cases and 3799 controls (including subjects from the initial study), identified 11 GW-significant SNPs at ten loci, verifying the previous findings and explained 0.647 (s.e.=0.048) MacTel heritability [132]. Out of all the genome-wide significant loci for MacTel, the strongest GW-association signal is driven by an independent rare risk haplotype, rs146953046 at *PHGDH* locus (1p12), with a frequency of 1.6% in controls and 5.9% in cases. Other follow-up studies from the 2017 GWAS demonstrated that low levels of serine in the blood serum of MacTel patients likely suppresses sphingolipid synthesis via serine palmitoyltransferase (*SPT*), which in turn enhances the accumulation of neurotoxic deoxy-sphingolipids, causing cytotoxicity [133], [134]. The latest GWAS also indicated the role of sphingolipids and cholesterol metabolism in MacTel by identifying three loci with genes (*CERS4*, *SUMF2* and *TTC39B*) overlapping in sphingolipid, as well as cholesterol metabolism. However, the mechanistic contribution of these metabolites in the progression of MacTel remains unclear. Recent studies to characterise the metabolomic profile of MacTel has demonstrated that dietary depletion of serine in a mouse study could result in a MacTel-like ocular phenotype [133] and also identified putative functional associations with glycerophospholipids as well as methionine–cysteine metabolic groups [134]. Additional GW-significant loci from the MacTel GWASs, encompass genes such as *TIMP3* (TIMP metallopeptidase inhibitor 3) and *SYN3* (synapsin III), and *SLC16A8*, *SLC6A20* (Solute Carrier family genes) which were also found to be associated with AMD [127], [132].

1.7.3 Transcriptomic studies of the retina

Only a few population-scale cohorts are available for studying healthy retinal tissue. Below three studies are described, comprising donor samples from the healthy retina, a tissue not part of the GTEx database.

Pinelli *et al.*, 2016

A study with 50 retinal RNA-seq samples from healthy retina [135] of post-mortem donors was performed to pioneer the *de novo* assembly of retinal transcripts as well as to quantify the gene expression of known transcripts, identifying 77,623 transcripts from 23,960 genes. Essentially, 92% of them were multi-exonic with 81% having known isoforms, 16% with unique isoforms and 3% were novel genes, determining retinal expression for 65% of the protein-coding genes in the human genome. The expression data also comprises more unusual gene biotypes such as small nuclear RNAs and miscellaneous RNAs.

Ratnapriya *et al.,* 2019

More recently, another transcriptomic study [94] with a cohort of 105 healthy and 348 AMD (total 453) post-mortem donor retinas was performed, which also assembled a retinal transcriptome, and performed eQTL analysis. This study established the Eye Genotype Expression (EyeGEx) database [94], mapping cis-eQTLs for the retina which comprised 14,565 genetic variants (eSNPs) that influence the expression of 10,474 genes (eGenes), becoming the first study to map the retinal eQTLs. Further, differential expression analysis of the retinal transcriptome was performed between the control and samples from AMD sufferers, adjusting for sex, batch effects analysed with or without age as a covariate. This analysis detected only 14 genes with, and 161 genes without correcting for age, at a false discovery rate (FDR) <= 0.2, identifying not many gene expression changes.

Strunz *et al.*, 2020

The latest mega-analysis of retinal eQTLs [136] was performed with a healthy cohort of 311 individuals by combining samples from three sources (including the control samples from Ratnapriya *et al.* study) making it the largest retinal eQTL study to date. 403,151 unique eSNPs, regulating 3,007 unique eGenes were identified, with more than one-third of the

regulated genes being long non-coding RNAs (lncRNAs). A regulatory cluster analysis was performed by filtering eSNPs regulating three or more genes and then combining those eSNPs < 1Mb apart within a chromosome into a genomic region resulting in 96 regulatory clusters. The authors claimed that 31 of these clusters contain 130 genes regulated by the same signal. A correlation analysis of the eVariants with GWAS significant variants (665) from 16 published genome-wide studies specific to twelve different eye disorders was performed, which identified 80 eGenes with the potential association and 65 of 665 GWAS variants hypothesized to regulate gene expression in retinal tissue. This study generated a valuable resource for further exploration of the mechanisms of retinal diseases. Table 1.2 compares the study details and the analyses performed in previous retinal transcriptomic studies.

Table 1.2: Comparison of recent retinal transcriptome studies denoting the cohort differences and the analyses performed by Pinelli *et al.*, Ratnapriya *et al.* and Strunz *et al.*

	Pinelli *et al.*	**Ratnapriya *et al.***	**Strunz *et al.* [Ratnapriya *et al.* controls]**
Cohort size	50 healthy donors	453 (105 controls + 348 AMD cases)	344 [105]
Average age	61.4	61	67.4
Percent females	56%	53%	42.76%
Visual impairment	NO	Mixed	NO
Ethnicity	European	Mixed	Mixed
Platform (RNAseq/Genotyping)	Illumina HiSeq / Illumina Infinium omni-2.5-8	Illumina HiSeq / UM_HUNT_Biobank v1.0 chip	Illumina HiSeq / custom HumanCoreExome BeadChip; Infinium OmniExpress-24 v1.2 BeadChip; UM_HUNT_Biobank v1.0 chip
Transcriptome assembly	✓	✓	×
Gene-set enrichment analysis	✓	✓	×
Gene co-expression network	✓	✓	✓
eQTL Analysis	×	✓	✓
Web interface for gene expression data	www.retina.tigem.it	-	-

1.8 Research scope and aims of this thesis

Understanding key drivers of gene regulation in the healthy and diseased retina remains a fundamental challenge in macular degeneration research, especially given the difficulty of accessing human retinal tissue. While numerous GWAS have identified promising disease loci, only a couple of studies to date have focused on the functional genomics aspects of the retina (despite the availability of retinal transcriptome [135]). eQTLs are widely used to interpret GWAS signals, to identify driver genes and biological mechanisms. Since transcriptomic data is tissue specific, eQTLs are often also tissue specific and hence identification of retina-specific eQTLs will assist in generating more mechanistic insights into complex retinal disorders. Therefore, in this thesis, I aim to perform the following analysis as elaborated upon further in the upcoming chapters:

- To identify cis-eQTLs from DNA genotyping (blood-based) and retinal transcript data in a novel cohort (Pinelli *et al.*) as an independent validation of healthy retina eQTLs;
    - compare and contrast the gene-expression and genotype data between the cohorts of Pinelli *et al.* and Ratnapriya *et al.*
    - explore power relationships for retinal eQTLs.

- Comparative analysis of the cis-eQTL results with the existing studies (aforementioned three datasets) to analyse the association between the eQTL signals.

- Application of eQTLs to understand functional genomics and provide mechanistic insights for the retinal disease *Macular Telangiectasia II* (MacTel).

Chapter 2: Description and data processing of a novel retinal eQTL dataset

2.1 Introduction

This chapter describes a novel retinal eQTL dataset generated by our collaborators from the Telethon Institute of Genetics and Medicine (TIGEM), Italy, who had published the RNAseq data previously in Pinelli *et al* [135]. The sample collection and data generation (gene expression and genotype data), as well as the quality control measures, we employed in the study, are described here. The input data requirements for a typical eQTL analysis are the genotype (DNA) and the expression data (RNA) for the study tissue. The recently available retinal eQTL results from Ratnapriya *et al.* [94] are used as a validation dataset.

2.2 Methods

2.2.1 Sample collection and data acquisition

Fifty post-mortem retinal samples from non-visually impaired individuals were collected by our collaborators in accordance with the tenets of the Declaration of Helsinki and after an informed consent from the donor or next of kin at Fondazione Banca degli Occhi del Veneto (FBOV). For each of the retinal tissue harvested, the metadata such as the age, gender, cause of death and the total post-mortem time (T) were recorded. The retinal tissues were isolated only from eye bulbs within a total post-mortem interval (T) <= 26 h (with mean T=20.5 h; range of T: 6 to 26 h) to reduce the possible post-mortem time effects on the RNA integrity and transcriptomic profiles. Donor tissue was considered to approximate the 'disease-free' retina, lacking obvious retinal pathology. Visual examination of the dissected retinae to exclude cross-contamination from adjacent tissues such as RPE/choroid was performed. The retinae were submerged in RNA stabilization Reagent (QIAGEN) immediately after dissection [135].

Our collaborators extracted RNA from the 50 samples using the miRNeasy Kit (QIAGEN) as per the manufacturer instructions. RNA quantity was assessed using a NanoDrop ND-8000 spectrophotometer (NanoDrop Technologies) and the RNA integrity

(average RNA integrity Number of 8.7) was obtained using an RNA 6000 Nanochip on a Bioanalyzer (Agilent Technologies). RNA libraries were prepared using the TruSeq RNA sample preparation kit and were sequenced on Illumina HiSeq 1000 platform via paired-end chemistry. For DNA genotyping, genomic DNA concentrations were evaluated using the High Sensitivity DNA Assay kit (Agilent Technologies) on a Bioanalyzer (Agilent Technologies) [135]. The retinal RNA expression data has been previously published [135]. The genotyping data is unpublished and analysed here for the first time.

2.2.2 Genotyping data and quality control

The genotyping data for the cohort (48 out of the 50 samples, as two samples were unsuccessfully genotyped) was generated by our collaborators using the Illumina Infinium Omni2.5-8 v1.4 BeadChip (Illumina, San Diego, CA) platform, containing 2,619,927 probes with a tagging ability of $r^2 >= 0.8$. The data analysis was performed using the computational tool, PLINK (version 1.9) [57], after formating the raw data to PLINK's specifications with the help of GenomeStudio (ver1.9, PLINK Input Report Plug-in v2.1.4). At first, analysis to identify discordant gender information, unusual heterozygosity and to infer excessive missing information was performed using the modules *--check-sex, --het,* and *--missing* in PLINK respectively. Subsequently, using the module *–indep-pairwise* and *--genome,* closely related individuals using Identity By Descent (IBD) were identified. This was performed by pruning the SNPs in the cohort for high linkage disequilibrium (LD) within a window of 50 base pairs with a step size of 5 and LD (r^2) = 0.2, between pairs of SNPs. This resulted in a coefficient representing the proportion of IBD (PI_HAT) and one of the individuals of the pair with PI_HAT > 0.185, was excluded from the analysis. Next, to infer sample ethnicity and identify any irregular clusters or outliers within the data, a principal component analysis (PCA) was carried out using the *--pca* module, including ~48,000 SNPs from the subjects and corresponding genotype information from the 1000 Genome Project reference data (Phase III).

The quality control (QC) [137] of genotype data included: retaining only autosomal variants, removing SNPs deviating from Hardy-Weinberg Equilibrium (HWE) (P-value < 1E-6), and retaining SNPs with Minor Allele Frequency (MAF) > 0.01 and a SNP genotyping

rate > 95%. Next, the quality-controlled SNPs were imputed to the Haplotype Reference Consortium reference panel (v.1.1) [36] via the Michigan Imputation Server [39] using Eagle (version 2.4) for phasing [138] and the minimac4 method for imputation [139]. The post-imputation quality assessment of SNPs was performed after converting them to chromosome-wise VCF format, using the following thresholds: imputation quality score > 0.3, MAF >1% and deviation from HWE P > 1E-6. The remaining high-quality SNPs were used for eQTL analysis (chapter 3) after their genomic positions were updated to hg38 using the UCSC *liftover* [140] -based R Bioconductor package, *liftOver* (version 1.8.0) [141].

2.2.3 Gene expression data and quality control

The transcriptome data for the Pinelli *et al.* cohort of fifty retinae from disease-free individuals was downloaded from the EBI ArrayExpress database (accession: E-MTAB-4377). As the obtained data was in bam format, the raw read data for the cohort was generated using the bamToFastq module from bedtools (version 2.26.0) [142]. The quality of the raw reads was inspected using FastQC (version 0.11.8). To remove the adapter and low-quality sequences, reads were trimmed using Trimmomatic (version 0.36) [143] in paired-end mode (SLIDING WINDOW 4:15; LEADING 3; TRAILING 3; MINLEN 30). The resulting paired-end reads were aligned to the Ensembl human reference genome (Genome Build Hg38, GRCh38.97) using STAR software (version 2.6.1) [41] with default parameters. The abundance of the aligned read counts were estimated for each reference gene product (considering the complete gene length obtained from a GTF file) using featureCounts [43]. The gene expression abundance was filtered by applying a threshold of CPM > 1, present in at least 10 % of the samples. The gene expression counts were transformed into log-normalised Counts Per Million (CPMs) using the R package *edgeR* (version 3.26.8) [144]. The gene expression data was converted into a chromosome-wise BED format with gene annotations for further analysis.

2.2.4 Identifying duplicate samples in genotype data

We observed some inconsistencies in the sample labels while performing the quality control analysis. In order to match the correct genotype sample to its corresponding sample in RNAseq data, and to remove duplicate samples, we obtained molecular evidence by

comparing the genotype calls from SNPchip with the variant calls from the RNAseq data of Pinelli *et al.* After processing the RNAseq data, as outlined above, the variants were called using the HaplotypeCaller module from the GATK tool kit using default parameters [145]. Since the SNP- chip genotype calls were based on human genome build Hg19, the variant calls from the RNAseq data were lifted-over using Picard tools (version 2.9.4) [146] to obtain uniform genomic coordinates for the SNPs from both technologies. Then, we compared each SNPchip sample against the entire RNAseq data cohort i.e., we attempted to match the samples by comparing their genetic variation. We encoded identical SNP genotypes as 1 (defined as an exact match [e.g., G/A= G/A], or a flip of the exact match [e.g., G/A=A/G] and either a complement [e.g., G/A=C/T] or reverse complement of the exact match [e.g., G/A=T/C], as RNAseq data was not strand-specific). All remaining cases were encoded as 0, or not matching. The proportion of exact matches for all pairwise comparisons was obtained and visualised.

2.3 Results

2.3.1 Genotyping data

A total of 41 samples remained after filtering low-quality samples out of the 48 genotyped libraries. A total of seven out of the 48 genotyped samples were excluded from further analysis due to low quality and two of the samples could not be genotyped out of the 50 subjects. Hence, we excluded seven of the samples due to quality control measures such as unusual heterozygote rate, extreme outliers of the PCA of the genotyping data and duplication as detailed below.

We excluded three samples due to unusual heterozygosity, defined as a heterozygosity rate at least 2 standard deviations from the mean (Figure 2.1, horizontal dashed lines). All samples were within the threshold of < 3% of missing genotypes [137]. This heterozygosity assessment was performed at an individual sample level. A population-level mean heterozygosity (mean 2pq component of Hardy-Weinberg equilibrium) estimation for the cohort was performed based on the minor allele frequency (MAF) and was compared to that of estimates for super-ethnicities from the 1000 Genome Project. This

helped to verify the original ethnicity (European) of the cohort (Figure 2.2). The cohort's ethnicity was further investigated for identifying population stratification, if any, by plotting the first two principal components. The Pinelli *et al.* samples clustered with the European (EUR) ethnic reference individuals, as expected from the meta-data of the cohort (Figure 2.3).

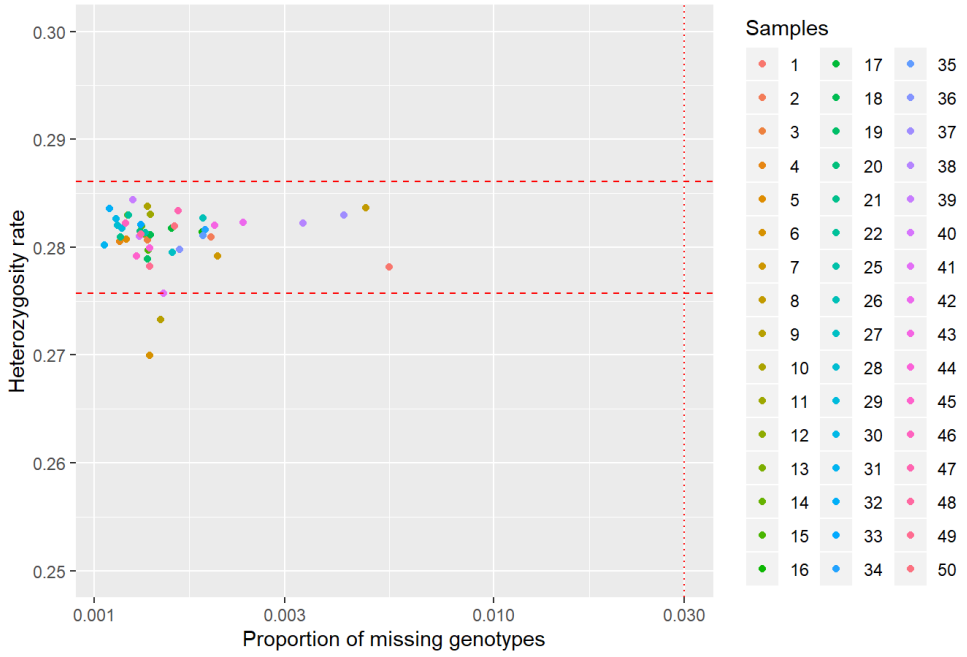

Figure 2.1: Estimation of the proportion of missingness (x-axis) and the heterozygosity of samples in the study cohort (y-axis). The samples (coloured points on the plot) 6,9,41, outside the horizontal dashed lines (heterozygosity rate ± 2 s.d.) were excluded from the analysis.

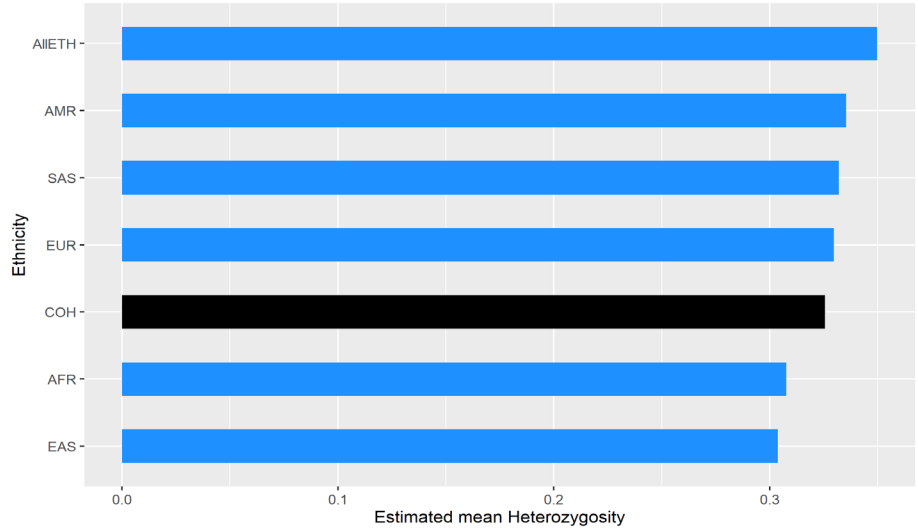

Figure 2.2: The estimated mean heterozygosity (x-axis) for Pinelli *et al.* cohort (y-axis black bar) and the super-ethnicities from Thousand Genome Project (y-axis, blue bars) are compared. The estimate is based on the MAF values of the individual and combined super ethnicities. The cohort's heterozygosity is close to that of the European population, as expected. ALLETH: combined super ethnicities; AMR: Ad Mixed American; SAS: South Asian; EUR: European; COH: Pinelli *et al.* cohort AFR: African; EAS: East Asian.

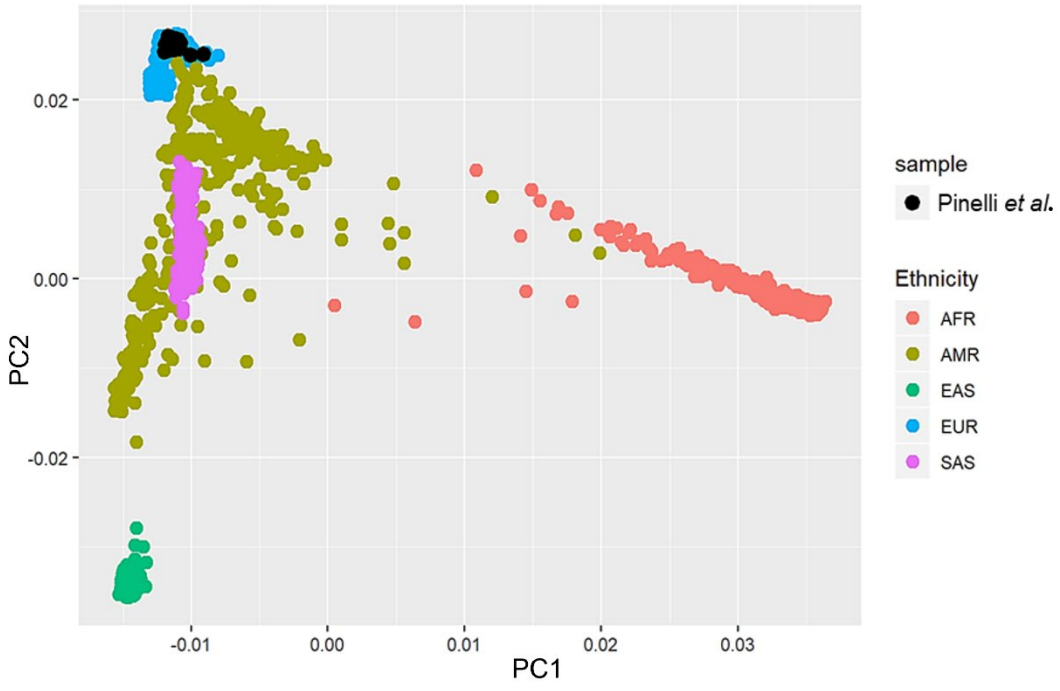

Figure 2.3: Ancestry clustering of the Pinelli *et al.* retinal cohort using the SNP chip data to verify subject ethnicity by merging with individual genotype data from the Thousand Genomes Project. AFR: African; AMR: Ad Mixed American; EAS: East Asian; EUR: European; SAS: South Asian; Pinelli *et al.*: Study sample.

After performing PCA analysis on the genotype data using the aforementioned SNPs (PCA performed with ~ 48,000 SNPs) for ethnicity identification, three more samples, being extreme outliers, were excluded from the analysis (Figure 2.4). 1,546,314 SNPs from the entire cohort that fell within the quality thresholds for MAF, heterozygosity, and missingness were retained. Finally, these high-quality SNPs after quality control were used to impute additional SNPs via the Michigan imputation server. After performing the post-imputation quality assessment as per the initial thresholds, 8,894,864 high-quality SNPs survived for the next stage of analysis.

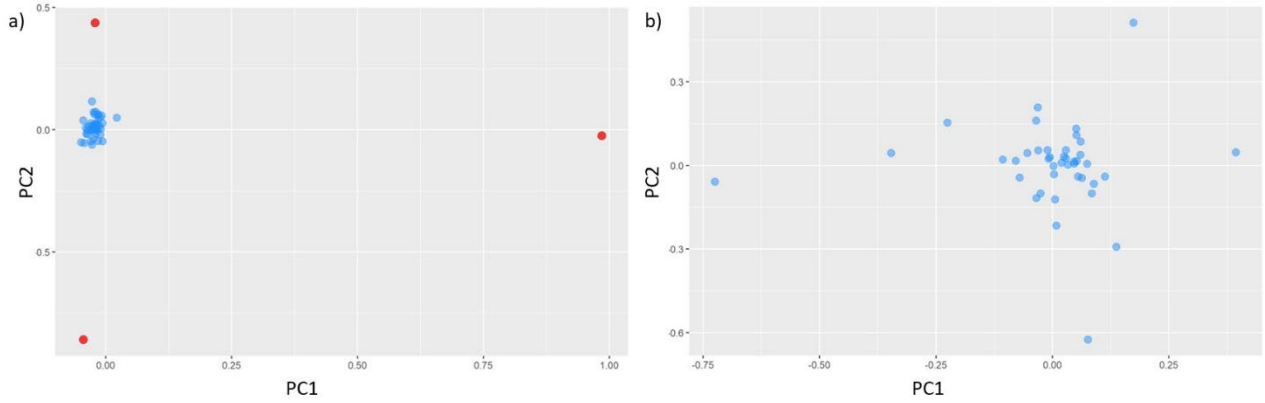

Figure 2.4: Principal Component Analysis of the samples with PC1 and PC2 on the x and y axes respectively. PCs a) before, b) after removing the outlier samples.

While comparing the genotypes with variants from RNAseq with a proportion of SNP-matching greater than 90%, sample S36 was excluded from the analysis after identifying a very close relationship with another subject (Figure 2.5). This subject was filtered out from the analysis because it likely represented a duplicate sample. This visualisation of this comparison helped in resolving the labelling issues.

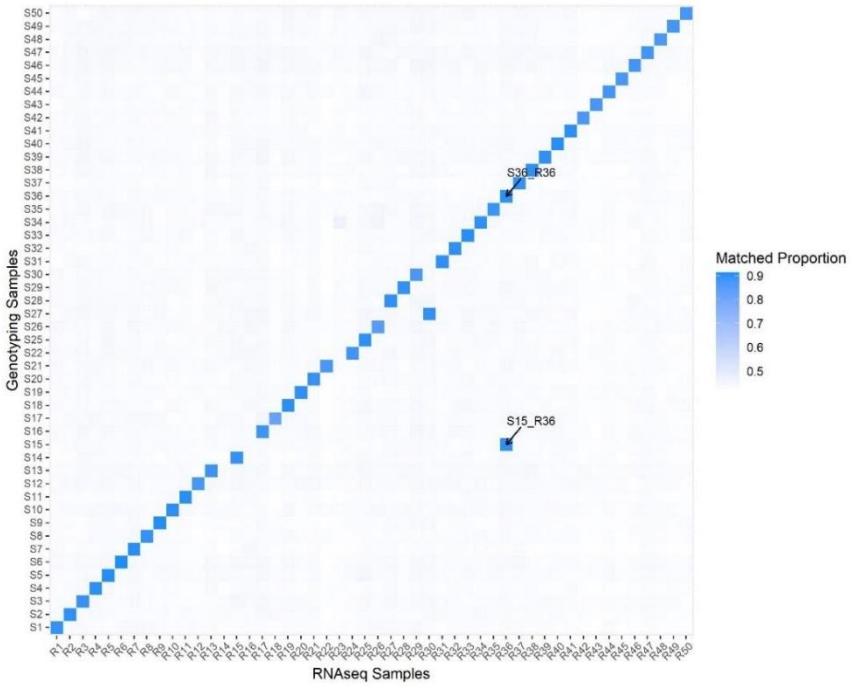

Figure 2.5: Comparison of SNPs genotyped and SNPs identified through RNA-seq reads between all pairs of samples. SNPs from both methods were matched and compared by a match proportion to identify the corresponding genotyping and RNAseq samples. One of the samples (labelled S15_R36) was a clear duplication of the samples 15 and 36 with a matched proportion of SNPs > 90%. Additionally, the IBD analysis through PLINK demonstrated higher missing genotype call rates for sample 36. Hence sample 36 was omitted from further analysis. X-axis: RNAseq data samples (labelled R1-R50); Y-axis: Genotyped samples (labelled S1-S50). Blue colour blocks represent the highest matching proportion of the SNPs.

2.3.2 Gene expression data

The median library size of the transcriptome data was ~ 28.9m reads. After performing the genome alignment and estimating the abundance of the gene expression, a threshold of CPM > 1 was applied resulting in a dataset with data from 22,794 genes. Nearly 14% of these genes were found to be non-coding RNAs. The Pinelli *et al.* data was compared to available data from the Ratnapriya *et al.* retinal expression study (processed with same RNA-seq workflow as applied to Pinelli *et al.*), and it was found that the Pinelli *et al.* dataset contained a sequencing yield nearly double that of Ratnapriya *et al.* (Figure 2.6). The gene expression abundance for the genes in common (12,584 genes) between the high-quality samples of the two cohorts were compared both individually and combined, with the cases and control samples of Ratnapriya *et al.* with the Pinelli *et al.* cohort. The two expression

data sets were found to be highly correlated with a correlation coefficient (Pearson correlation, R) = 0.82 (Figure 2.7c).

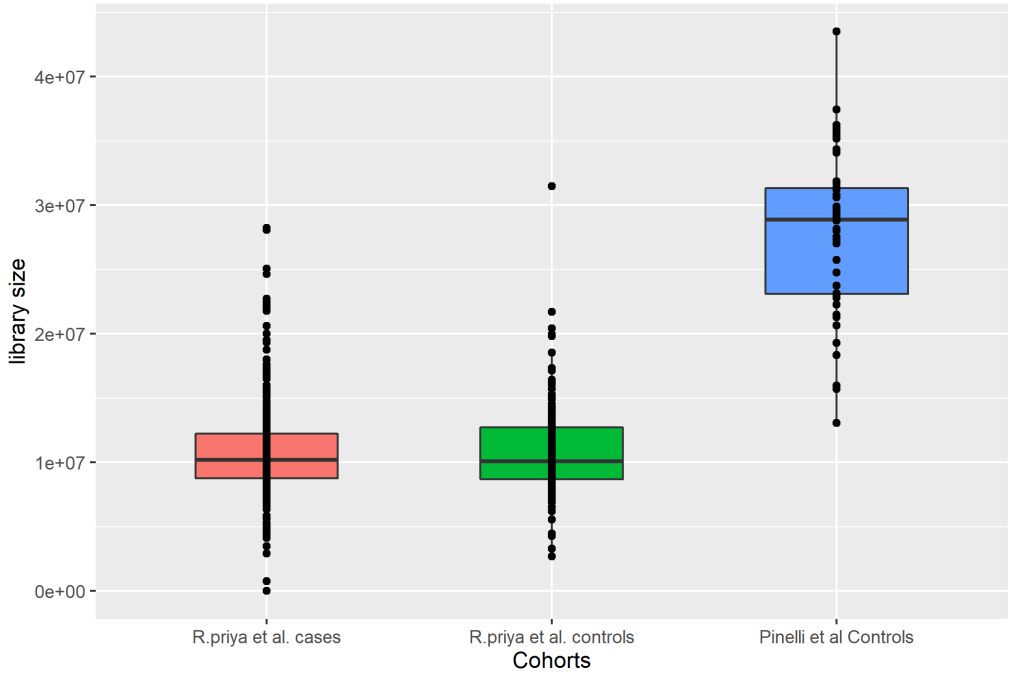

Figure 2.6: Read library sizes of AMD cases and controls from Ratnapriya *et al.*, and Pinelli *et al.*

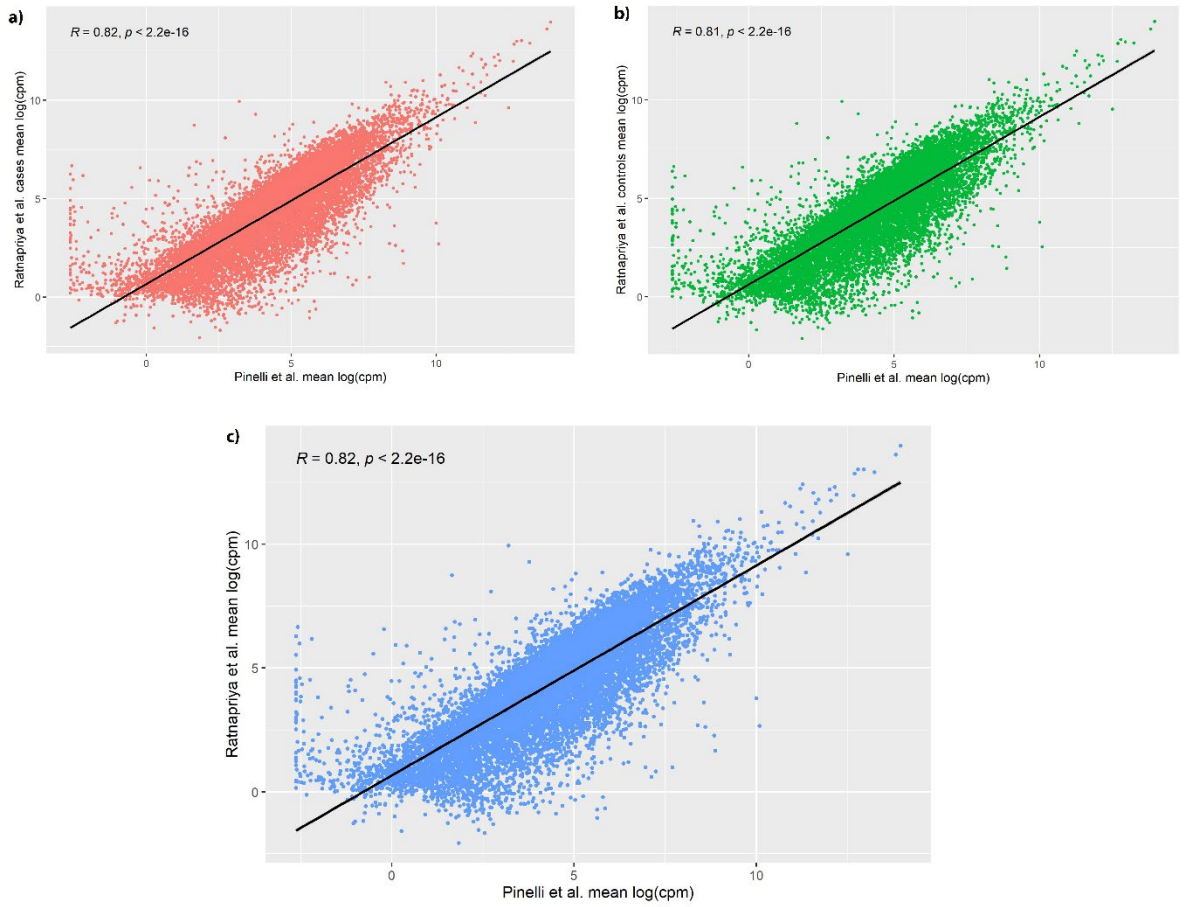

Figure 2.7: Comparison of gene expression data between the high-quality samples for the cohorts. a) Pinelli *et al.* and Ratnapriya *et al.* cases b) Pinelli *et al.* vs. Ratnapriya *et al.* controls. c) All samples in both cohorts. The points on the plot represent the average log-CPM value for the common genes (12,584) between the data sets. The black line represents the y=x line or equal abundance.

2.4 Summary

We completed the quality control analysis of the genotyping data from the same cohort (Pinelli *et al.*) of individuals that had previously undergone expression analysis generating the first-ever retinal transcriptome [135]. A snapshot of the data set under study is given in the table below (Table 2.1). We lost seven of the genotyped samples after quality control making the sample size even smaller. Notably, this data set is very homogenous from the ancestry perspective, which is identified as European. Also, the depth of the gene

expression (RNA-seq) data is almost double compared to a recent retinal eQTL study, which provides greater power to examine low abundance genes. Our dataset has expression data on ~30% more genes than Ratnapriya et al, with greater read lengths. However, the modest sample size in our dataset is a limiting factor for adequate statistical significance in the downstream analysis for eQTL. Our dataset is only ~10% of the size of the very large Ratnapriya et al dataset. With the completed quality control steps of the expression data and the genotype data, it is now possible to perform an eQTL analysis of the Pinelli *et al.* data.

Table 2.1: Summary of the Pinelli *et al.* and Ratnapriya *et al.* datasets.

Data set	**Pinelli *et al.* cohort**	**Ratnapriya *et al.* cohort**
Genotyping Platform	Illumina Infinium Omni2.5-8 BeadChip	UM_HUNT_Biobank v1.0 chip
Number of Samples	48	516
Samples remained after Quality Control (QC)	41	406
SNPs surviving QC for imputation	1,546,314	570,441
SNPs surviving imputation QC	8,894,864	8,924,684
Tool & Reference for Phasing and Imputation	Michigan Imputation Server + HRC panel	IMPUTE2 + Thousand Genome Project panel
RNA-Seq library	TruSeq RNA Library Prep Kit	TruSeq RNA Library Prep Kit
RNA-Seq Sequencing platform	Illumina HiSeq 1000 platform	Illumina HiSeq 2500 platform
RNA-Seq Sequencing depth	~28.9m PE	~10m PE
Read length	150bp	125bp
Expressed genes (CPM > 1 in 10% of samples)	22,794 genes	17,374

Chapter 3: eQTL analysis of a novel retinal dataset

3.1 Introduction

The previous chapter explained the genotype and gene expression data processing, and quality control steps. A total of 41 samples out of the 48 samples with RNA and genotype data remained for the eQTL analysis after quality control. QTLtools [77] was previously determined to be a fast and efficient permutation-based eQTL identifier with a versatile modular framework for each step of the analysis, and was used in this eQTL analysis. The permutation analysis identifies the top eQTL for a gene with fewer permutations by estimating the beta distribution of the P-values. QTLtools was also used by the recent retinal eQTL study, Ratnapriya *et al.* [94]. In contrast, its preceding version, FastQTL [88], was employed for analysis of the GTEx consortium [68], [147] data. Using QTLtools, our preferred tool for identifying eQTLs in the retinal tissue, the genotyping and phenotyping data from Pinelli *et al.* were integrated using linear models in pursuit of retinal eQTLs. The steps followed for the eQTL analysis are summarised in a flow diagram (Figure 3.1).

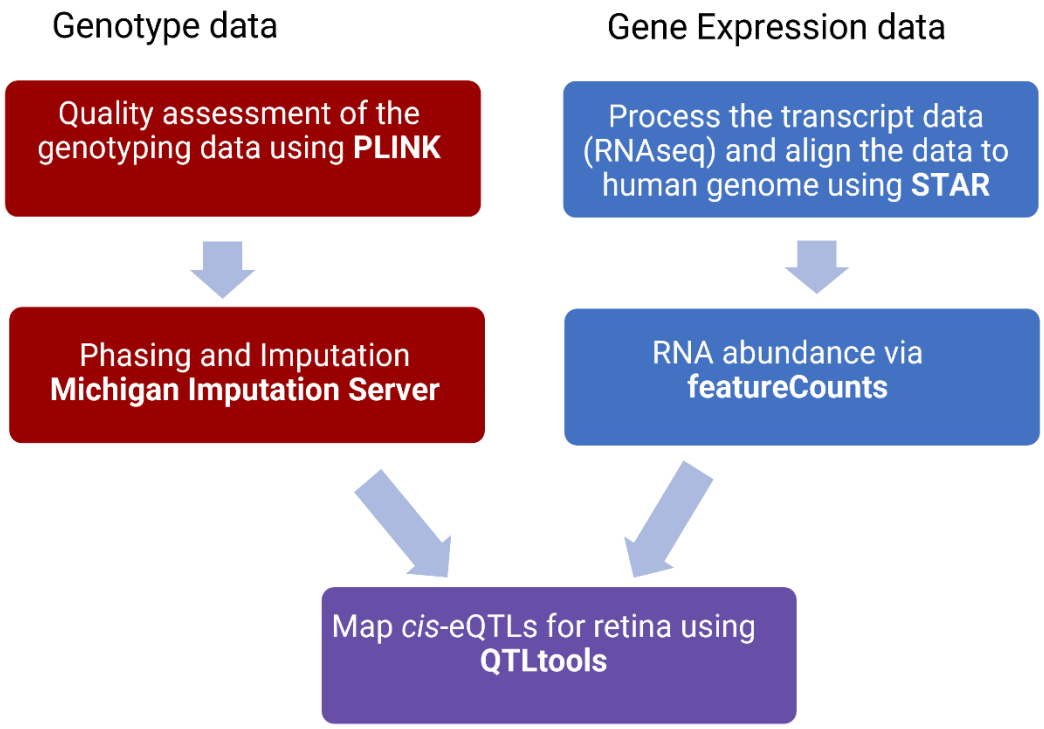

Figure 3.1: Workflow designed for the eQTL analysis. Red= DNA genotype data; Blue= RNA transcript data.

3.2 Methods

3.2.1 Pilot data analysis

A pilot dataset for the eQTL analysis is provided along with the QTLtools package, containing 358 samples of European origin. The genotype and gene-expression (RNA-seq) data for these samples is derived from the 1000 Genome Project and the Geuvadis project respectively [77]. At first, we validated the input requirements for QTLtools by comparing our inputs with this pilot dataset. We identified that the normalisation of the RNA-seq data required for the eQTL analysis was estimated in terms of Reads Per Kilobase per Million (RPKM) mapped reads. Therefore, we implemented the normalisation of the gene-expression data to RPKM and proceeded with the analysis.

Previously published retina eQTL datasets

The retinal eQTL summary statistics for the previously published studies were downloaded from the links provided in the publications. These studies were described in section 1.7 of the literature review (chapter 1). The results of the first retinal eQTL study [94], henceforth referenced as the Ratnapriya *et al.* study, were downloaded from the GTEx portal (https://www.gtexportal.org), which is available as an external resource to the original GTEx data. The retinal mega-analysis study [136] results, henceforth referenced as the Strunz *et al.* study, were accessed from the website, https://www-huge.uni-regensburg.de, as provided by the authors in their publication. The summary results from these two studies were used to perform the comparative analysis of retinal eQTLs.

3.2.2 eQTL analysis with QTLtools

The eQTL analysis for this dataset was performed employing the *--cis* module of QTLtools (Version 1.3.1), which uses an additive linear regression model to identify proximal eQTLs. The genotype (VCF format) and RNA-seq (RPKM > 1 in 10% of samples; in bed format) data were given as input as prescribed in the QTLtools manual. For each gene,

the *cis*-eQTLs were identified within a window of 1 Mb upstream and downstream of the transcription start site. The coordinates for genes were annotated as defined by Ensembl [148]. Gender, age and one genotype PC (as this explained the maximum unwanted genetic variation in the data) were included in the linear regression model as covariates. At first, the most significant associations between a gene (eGene) and its nearby SNP (eSNP) were identified by using the option *--permute 1000* along with other recommended parameters of the QTLtools. This step resulted in a beta distribution extrapolated empirical P-values for each eGene:eSNP pair. Next, these P-values were corrected using the false discovery rate (FDR) method from Storey *et al.* [149] to calculate the Q-value for each gene using the *qvalue* (version 2.18.0) R package. The most strongly associated eSNP for each eGene was identified by applying a q-value threshold of 0.1. Further, as described by the GTEx consortium [91], to identify all independent significant eQTLs for an eGene, a nominal P-value threshold for each gene was calculated using the previously determined beta distribution parameters of QTLtools. The FDR correction and threshold estimations were applied using the R script, *qtltools_runFDR_cis.R*, provided along with the QTLtools package. Thereafter, QTLtools was run in nominal mode with recommended parameters to obtain the nominal p-values for identifying all secondary independent eSNPs for an eGene. The eSNPs were considered significant if the nominal P-values were below the gene-specific nominal P-value threshold determined using the beta distribution parameters in the earlier step. The eQTL slopes obtained for the matching top eGene:eSNP pairs were compared with recently published retinal eQTL studies.

3.2.3 Power analysis

The power analysis for this cohort was initially performed using the state-of-the-art RShiny application, powerEQTL, hosted at https://bwhbioinfo.shinyapps.io/powerEQTL. In this web application, we opted for simple linear regression as the statistical model amongst the other models (e.g., ANOVA) available, the number of tests as 500,000 and the sample sizes of 41, 100, 150, 200. However, for estimating the power over a range of inputs simultaneously *viz.* various MAFs, number of tests and detectable slopes, we used the available R package of powerEQTL. Using this package, we estimated the power of an eQTL

study given the number of tests, MAF threshold and the expected minimum effect size. The function *powerEQTL.SLR()* was supplied with the number of SNPs as the number of tests, a range of MAF values from 0.05 to 0.4 and the minimum detectable absolute slopes ranging from 0.5 to 2.5 as input for power estimation.

3.2.4 Technical exploration of statistical significance and power in the Pinelli et al. data

Restricted eQTL analysis

The insights from the initial eQTL analysis as well as power analysis for this cohort indicated the need to increase power through mitigation strategies such as a reduction of the multiple-testing correction burden. Therefore, we performed eQTL analysis with restricted inputs to reduce the multiple testing burden by reducing the input genotype as well as the number of genes (see "Restricted gene set analysis"). At first, the genotype data input for QTLtools was filtered to reduce the number of statistical tests and thereby anticipated an increase in statistical significance after controlling the FDR. The genotype data filtering was achieved by applying a range of MAF thresholds from 0.05 to 0.4 and the resulting restricted data was used as input to run QTLtools. The aforementioned FDR correction procedure was followed for each run.

Restricted gene set analysis

For AMD and MacTel genes

A gene set specific to retinal disorders was obtained by selecting genes associated with AMD and MacTel, from the large AMD GWAS [124] and the latest MacTel GWAS [132] respectively. This set contained a total of 851 genes: including 54 MacTel genes with association to at least one suggestively significant risk loci, and 797 candidate genes from the 34 AMD associated loci determined by the 52 GWA-significant variants and their proxies (r2 $\geq$0.5, $\pm$500 kb) published by Fritsche *et al.* At first, we checked if these genes were expressed in the Pinelli *et al.* data. The expressed genes were matched to the top eGene:eSNP

results. An FDR correction at thresholds of 0.05, 0.1, 0.2 and 0.3, was performed on the matched set to identify the significant genes from this restricted gene set.

For highly expressed genes

We also performed a restricted analysis to check the significant eQTLs in the highly expressed genes by ranking them based on expression and variance. We ranked the genes from the top 10% to 50% highly expressed genes along with calculating the coefficient of variation (CV) (S.D. of gene expression / mean of gene expression). The CV was used to visualise the proportion of genes at each ranking threshold of high expression. Each set of genes, at various ranking thresholds, was tested for statistical significance (FDR) individually as well as combined with the retinal disorders specific gene set selected in the previous step.

Comparison of effect-sizes for MacTel-specific eQTLs

An analysis to investigate the agreement in effect size and direction for the eQTLs specific to MacTel was performed. This compared the eQTL effect sizes (a.k.a 'slopes' or 'beta values') from the Pinelli *et al.* data with the MacTel-specific eQTLs identified using the EyeGEx data [94] in the recent MacTel GWAS by Bonelli *et al.*, 2021 [132]. The nominal results of the Pinelli *et al.* data were matched to Ratnapriya *et al.* eQTL results, and concordant (e.g. positive slopes for matching eSNP:eGene pairs in both results) and discordant (e.g. opposite direction of slopes for matching eSNP:eGene pairs in both results) eQTLs were counted to populate a Punnett square. An unweighted Cohen's kappa statistic was determined to obtain the level of concordance for the direction of effects, between the matched MacTel-related eQTLs obtained previously from Ratnapriya *et al.* with those from Pinelli *et al.* and Strunz *et al.* results, using the *psych* [150] R package.

3.3 Results

3.3.1 eQTL analysis

After performing quality control, the cohort size for performing eQTL analysis was 41 samples. We performed a local eQTL analysis to detect *cis*-eSNPs within a maximum of

1Mb genomic distance either up- or down-stream of each gene transcription start site. Four genes were found to be statistically significant after applying an FDR threshold of 0.1, with 333 independent unique eSNPs affecting their expression. A snapshot of the results containing the top eQTL per eGene for this eQTL analysis is provided in Figure 3.2. It shows a Manhattan-style plot of the transformed P-values (Figure 3.2a), and the effect sizes obtained for each eSNP (Figure 3.2b), with the significant results coloured in red. The effect sizes ranged between ± 2.5. We observed a density peak of eSNPs near the TSS as expected (3.2c). Two of the four significant eGenes, *AC091729.3* (7p22.3) and *LINC01535* (19q13.12) are long non-coding RNAs. The remaining two significant protein-coding eGenes were *HSD17B12* (11p11.2) and *WFDC10B* (20q13.12). After examining these genes in the published retinal eQTL studies (Ratnapriya *et al.* and Strunz *et al.*), we found that all four genes are significant eGenes in both studies. The summary statistics for these four genes along with the gene-wise P-value threshold for secondary eQTL signals are presented in table 3.1, along with the genotype-specific expression for these eGene:eSNP pairs in Figure 3.3.

Considering the entire data, we observed that the density of the eSNPs was greatest towards the transcription sites of the eGenes (Figure 3.2c), consistent with the expectation of a regulatory SNP (eSNP) to be near an eGene. We also observed a loss of continuity in extreme beta values, which plateaued around ±2.5.

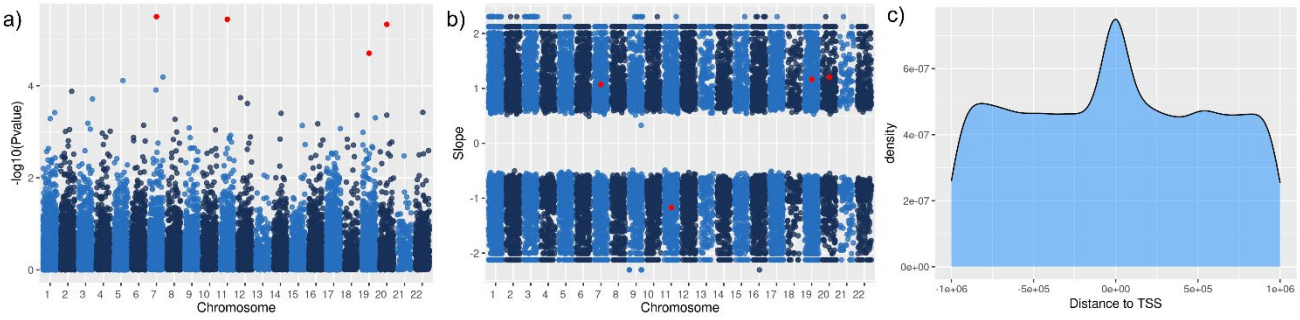

Figure 3.2: A snapshot of the results of the top eQTL per eGene for the Pinelli *et al.* cohort. a) Manhattan plot of P-values (y-axis) b) Equivalent plot of effect-sizes (beta; y-axis) c) The distance (x-axis) of the eSNPs from the transcription site of the corresponding eGene.

Table 3.1: The summary results of the four significant top eGene:eSNP pairs obtained at an FDR significance of 0.1.

ENSEMBL Gene ID	Gene Symbol	Chr	Gene start	Gene end	Distance of variant from TSS	SNP ID	Nominal P-value	Effect size	Adjusted P-value	Nominal P-value significance Threshold
ENSG00000229043	*AC091729.3*	chr7	1160375	1165267	-3346	7:1196665:C:G	4.232E-09	1.07265	3.161E-06	1.588E-08
ENSG00000149084	*HSD17B12*	chr11	43680559	43856617	0	11:43844500:C:T	5.166E-09	-1.1712	3.613E-06	1.710E-08
ENSG00000226686	*LINC01535*	chr19	37251913	37265535	-4965	19:37737850:A:G	4.967E-08	1.16468	1.968E-05	3.038E-08
ENSG00000182931	*WFDC10B*	chr20	45684654	45705019	-17763	20:44351421:C:A	2.234E-08	1.20493	4.648E-06	5.334E-08

Chr=Chromosome; Adjusted P-value = Adjusted empirical p-value obtained by the fitted beta distribution

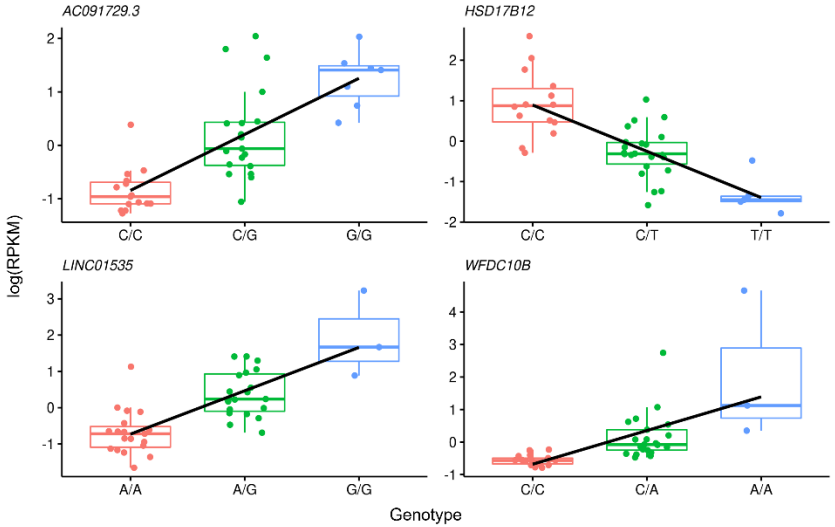

Figure 3.3: The allele-specific expression of the four significant genes from the Pinelli *et al.*, data at an FDR correction threshold of 0.1. The x-axis represents the genotype alleles, and the y-axis represents the log transformed RPKMs of gene expression.

3.3.2 Comparing Pinelli et al. eQTLs with recently published retinal studies

The results from the eQTL analysis after FDR correction (threshold of 0.05) were compared to the latest retinal eQTL studies, as shown in Table 3.2 below. In the first published study of adult retinal eQTLs, Ratnapriya *et al.* reported a greater number of significant results with a 10-fold larger sample size of 406 subjects. However, a mega-analysis study by Strunz *et al.* [136], that encompasses the control samples from Ratnapriya *et al.*, did not obtain as many significant results as Ratnapriya *et al.* despite the greater sample size. We compared the effect-sizes between the three cohorts for the eGene:eSNP pairs in common between the datasets (refer to Figure 3.8; presented towards the end of the section 3.3.4.), and found a poor correlation of effect-sizes between Pinelli *et al.* with the other two published eQTL datasets. There is a better correlation of the effect-sizes between the Ratnapriya *et al.* and Strunz *et al.* eQTLs, when compared to Pinelli *et al.*, not only due to the increased sample sizes but also since the first two studies are not independent, with Strunz *et al.* including the control samples of Ratnapriya *et al.*

Table 3.2: Comparison of the significant results from Pinelli *et al.* with published retinal eQTL studies at an FDR threshold of 0.05. Note: The published studies report results at an FDR < 0.05, and hence we compared the results at this FDR threshold.

	Significant top eGene:eSNPs		
FDR threshold	Pinelli *et al.* (n=41)	Ratnapriya *et al.* (n=406)	Strunz *et al.* (n=311)
0.05	3	10,474	3,007

3.3.3 Power analysis

Given the low eQTL signal in data from Pinelli *et al.*, we performed power studies for the data of Pinelli *et al.* using the interactive R shiny web application as well as the manual

powerEQTL [151] R package. The input values included a sample size (n) of 41 representing our cohort, along with arbitrary sample sizes such as 100, 150, 200 for comparison. Unfortunately, the sample size of 41 was unable to reach the conventional statistical power of 80%, even at a high minor allele frequency, suggesting that this is an underpowered sample for obtaining statistically significant results (Figure 3.4a). Although the RShiny application is good for estimating power for a fixed number of tests, the standalone R package is used to scale the analysis for estimating power for multiple tests simultaneously, at various MAF thresholds. We observed that the sample set reached a power of 80% only at high effect sizes and with large input data at a MAF threshold of ~0.25 for a moderate effect size (Figure 3.4b). The power studies demonstrated that the sample size was the dominant limiting factor in achieving significant power. Strategies such as reducing the number of genes and placing bounds on the MAF were not sufficient. This is consistent with other recent work [89]. Our power estimation for eQTLs, based on real genome data, corroborates the previous simulation-based eQTL study by Huang *et al.* [89], which suggested that eQTL studies with less than 100 samples are underpowered even to attain a moderate effect size.

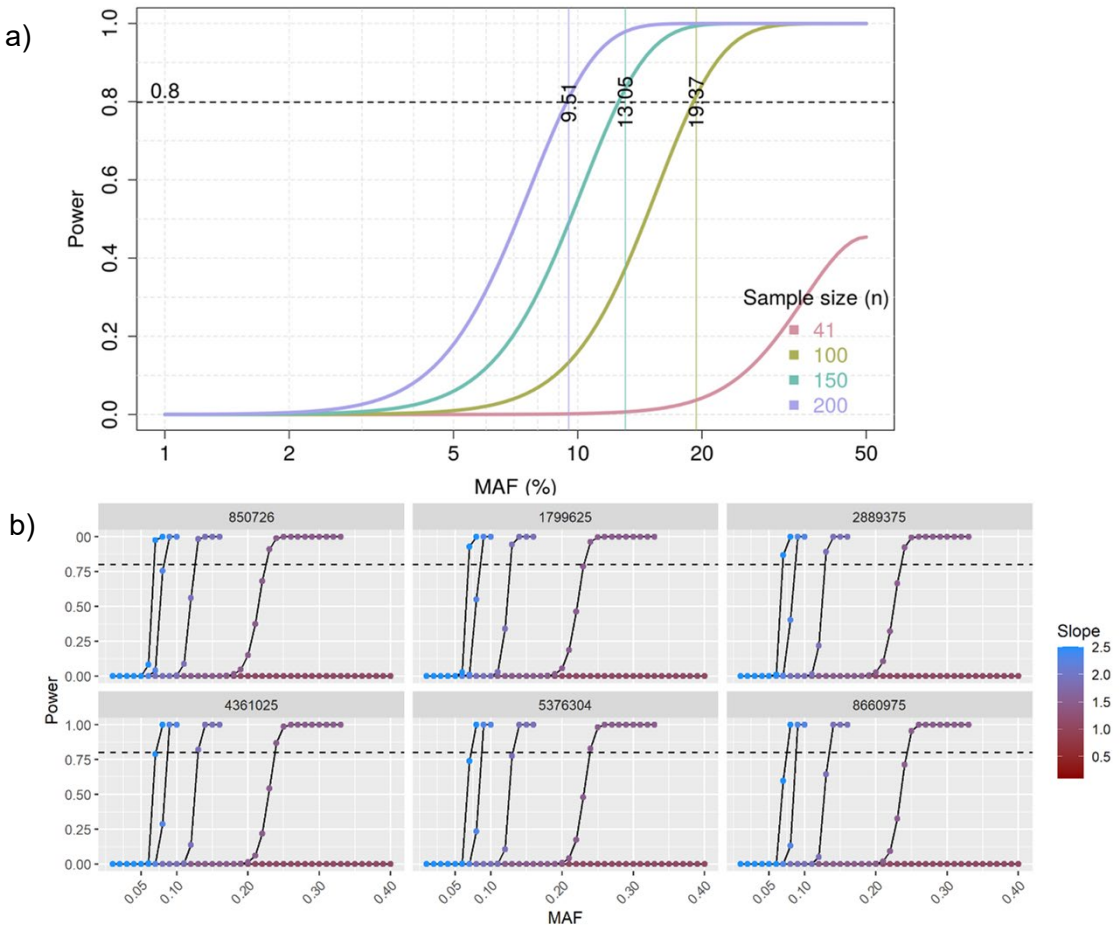

Figure 3.4: Power calculations for Pinelli *et al.* data using the powerEQTL. a) Power estimated (y-axis) through the RShiny application of powerEQTL with 500,000 tests for sample sizes of 41, 100, 150, 200 and MAF values (x-axis) ranging from 0.1 to 0.4. b) Power (y-axis) calculated using the powerEQTL R package, with SNPs at MAF defined values (x-axis) of 0.05, 0.1, 0.2, 0.3, 0.4 as the number of tests, for a minimum detectable effect-size (slope) ranging from 0.5 to 2.5. The black dashed line represents the 80% power threshold considered for statistically significant results. The facet heading numbers represent the number of tests given for each MAF threshold.

3.3.4 Exploration of the eQTL results

Restricted eQTL analysis

Analysis with the Pinelli *et al.* cohort suggested a weak eQTL signal driven by the small sample size, which was not mitigated by the deeper RNAseq libraries. Therefore, to further explore this data, with anticipation of a better statistical significance, we considered reducing the number of tests to reduce the multiple testing burden. With this rationale, we first applied independent eQTL runs with increasing MAF thresholds. The number of SNPs

that remained at each MAF threshold is shown in Figure 3.5a. We compared the p-values for minor allele frequencies from 0.05 to 0.4 (Figure 3.5b). We could observe that the log-transformed adjusted P-values were reduced (shifted right) for the eSNP subsets filtered at increasing MAF thresholds, indicating that these thresholds influenced the corrected p-value distribution in the expected direction. However, we ultimately found no increase in statistically significant results after FDR correction. This may be due to the fact that even though the number of eSNPs was reduced, this did not have much impact on the number of eGenes since many eSNPs contribute to an eGene. Therefore, we next attempted to reduce the number of genes being examined.

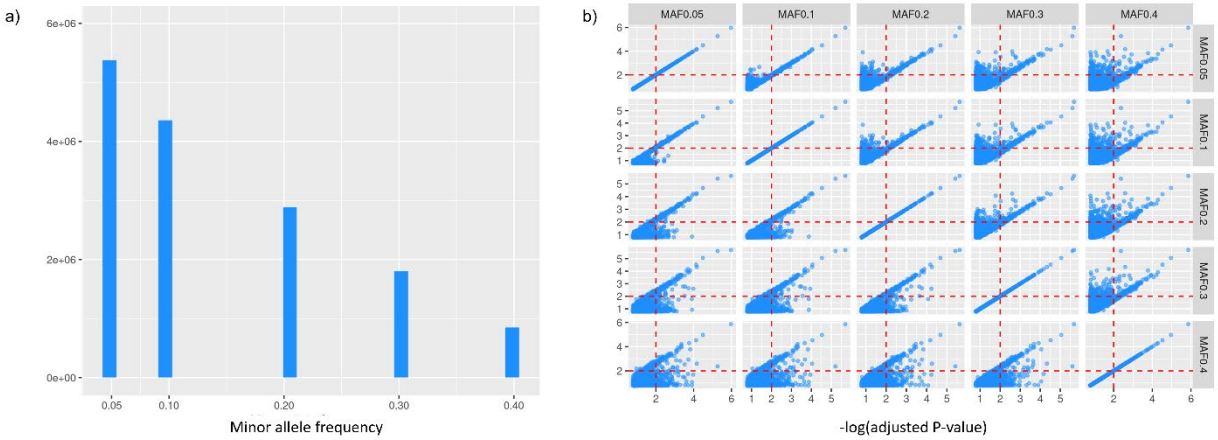

Figure 3.5: Analysis of relationship between MAF and significance for Pinelli *et al.* eQTLs. a) Number of SNPs (y-axis) remained at each MAF threshold (x-axis). b) Comparing the -log beta adjusted P values (blue points) obtained for the top eGene:eSNP pairs for various runs of QTLtools, with increasing MAF thresholds for Pinelli *et al.* data. We can observe the reducing trend of P-values (increasing points to the top/right of the red dashed lines) as the MAF threshold increases.

Highly expressed and disease-related gene set analysis

As we could identify only a few genome-wide statistically significant eQTL results for the Pinelli *et al.* cohort, we performed several restricted gene set analyses. First, we selected 851 genes that are associated with AMD and MacTel, obtained from Fritsche *et al.* 2016 and Bonelli *et al.* 2021 respectively, to gain further insight into the molecular biology of these diseases. Upon matching these genes with the Pinelli *et al.* RNAseq abundance, 691 of the 851 genes were found to have a mean RPKM > 1 and were thus considered as being

expressed in the retina (Figure 3.6a). Four and seven of the 691 expressed genes were found to be statistically significant top eGene:eSNP pairs using the FDR thresholds of 0.05 and 0.2 respectively (Table 3.3). All these seven genes were associated with AMD. The genotype-specific expression plots for the significant eGene:eSNP pairs were generated (Figure 3.6b) and as observed the majority of these eSNPs had an increasing effect on the gene expression with the minor allele coded as the effect allele. The genes *HLA-DQB2*, *HLA-DRB5*, *NOS2* were found to be significant eGenes in both Ratnapriya *et al.* and Strunz *et al.*, while the gene *DNTTIP1* is a significant eGene in Strunz *et al.*

Table 3.3: Significant results after the FDR correction on the restricted gene set associated with AMD and MacTel disorders.

ENSEMBL Gene ID	Gene symbol	SNP ID	Rsid	Description	Association with disease	FDR threshold
ENSG00000232629	*HLA-DQB2*	6:32605525:C:T	rs9272454	major histocompatibility complex, class II, DQ beta 2	AMD, Allergy	0.05
ENSG00000180083	*WFDC11*	20:44332298:T:G	rs432448	WAP four-disulfide core domain 11	AMD, Epididymis disease	
ENSG00000182931	*WFDC10B*	20:44401630:T:A	rs6065886	WAP four-disulfide core domain 10B	AMD, Testicular Leukemia	
ENSG00000168634	*WFDC13*	20:44385389:A:G	rs1711203	WAP four-disulfide core domain 13	AMD, Epididymis disease	
ENSG00000198502	*HLA-DRB5*	6:32546795:T:A	rs1064701	major histocompatibility complex, class II, DR beta 5	AMD, Multiple Sclerosis, Trochlear Nerve disease	0.2
ENSG00000007171	*NOS2*	17:26225558:G:C	rs1034895	nitric oxide synthase 2	AMD, Malaria, and Meningioma	
ENSG00000101457	*DNTTIP1*	20:43432280:C:T	rs6103872	deoxynucleotidyl transferase terminal interacting protein 1	AMD, Lymph Node Carcinoma	

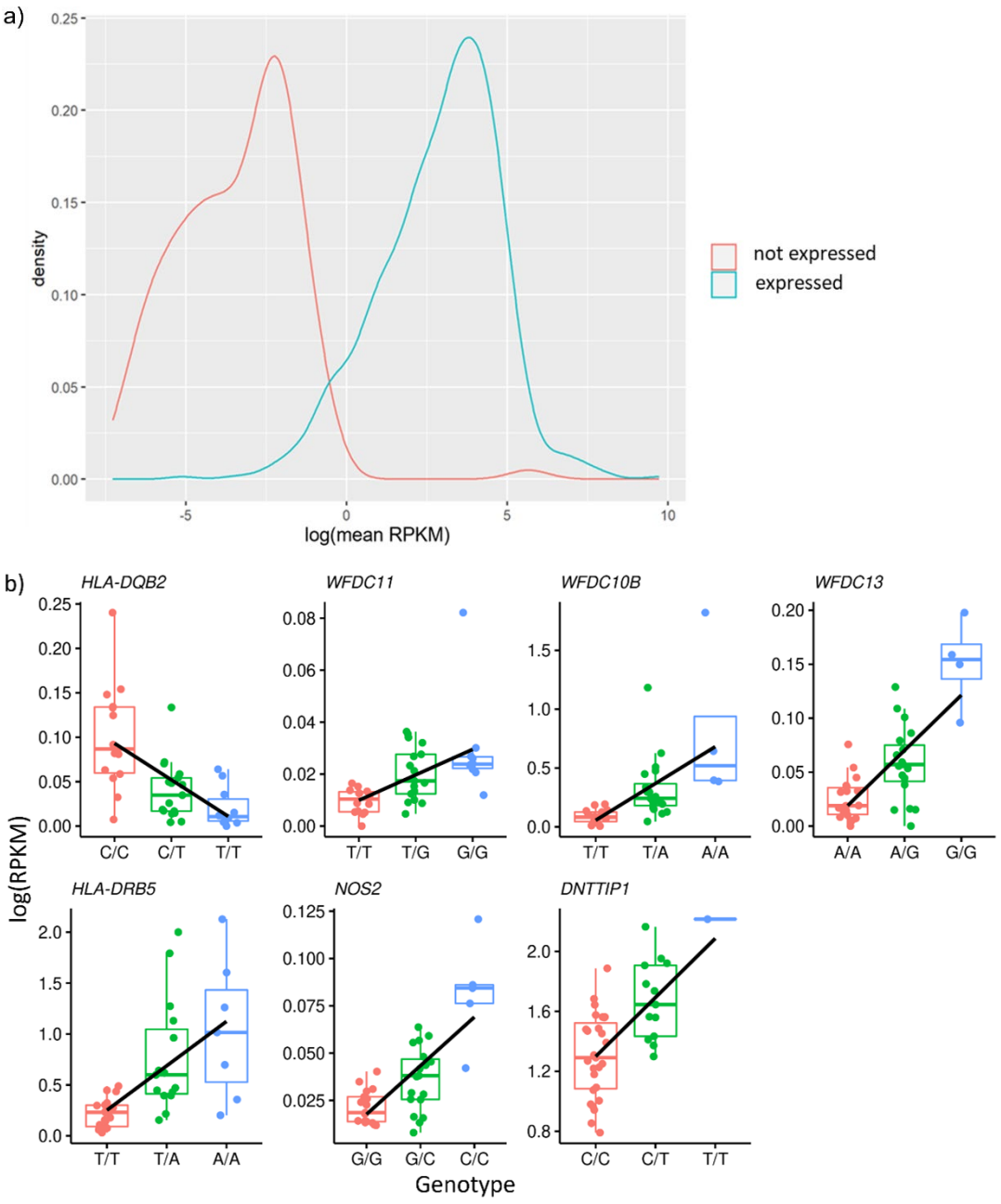

Figure 3.6: a) Examining the specific gene set (851 genes) for their expression in the Pinelli *et al.* data. 691 of the 851 genes were expressed at a mean RPKM > 1. The red and green density plots represent the genes not expressed and expressed in the cohort respectively. **b) The allele-specific expression of the seven significant genes, from the restricted genes expressed in Pinelli *et al.*, at an FDR correction threshold of 0.2.** The x-axis represents the genotype alleles, and the y-axis represents the log transformed RPKMs of gene expression.

Genes were ranked by expression values and an FDR correction was performed on a subset of genes of increasing expression rank to check whether a reduction in the number of genes increased the number of genes surviving the less severe multiple testing correction. The coefficient of variation for the genes reduced as the mean expression increased and hence the eQTL signal might be further reduced. The highly expressed genes for each decile of rank with their coefficient of variance were visualised in a mean-variance plot (Figure 3.7). We observed similar numbers of significant genes after FDR correction by merging the retinal disorder-specific genes (analysed in the previous step) to each subset of genes sampled by expression rank decile. All the significant genes (at FDR 0.2) were of the subset associated with AMD.

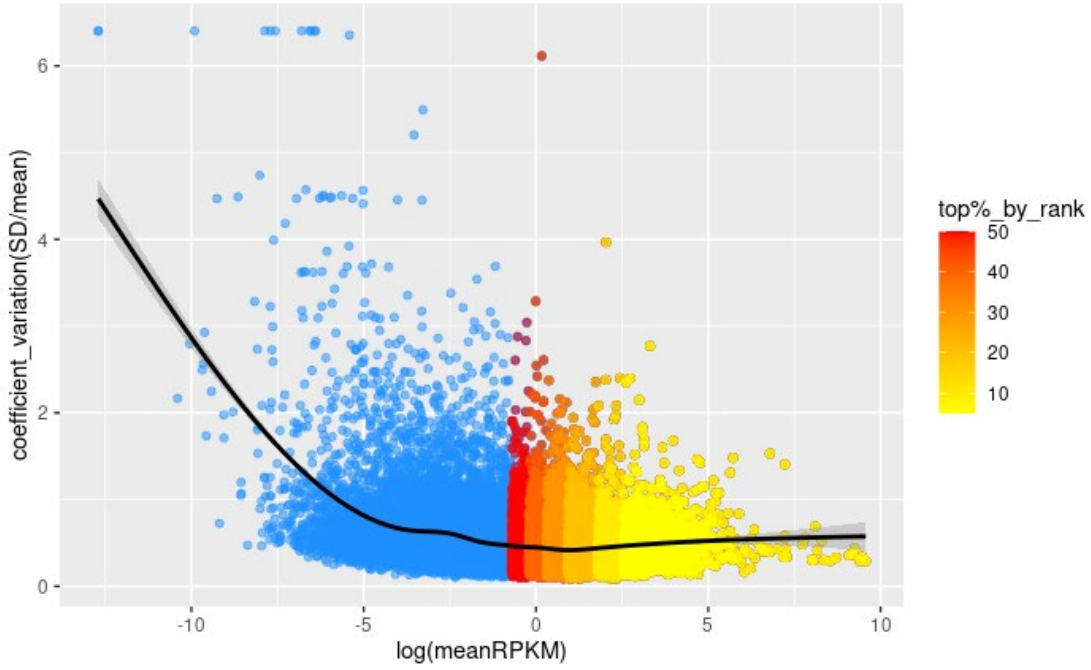

Figure 3.7: Ranking the Pinelli *et al.* genes based on expression and comparing their coefficient of variance. All the genes (blue) with RPKM > 1 are represented with the ranked deciles for genes (red colour gradient). The mean-variance of the gene expression (black line) is highly reduced as the gene expression increases.

MacTel-related eQTL analysis

Lastly, we were interested to observe the concordance of the effect size and direction of eQTLs between the EyeGEx [94] data and Pinelli *et al.* within disease risk loci for the recent genome-wide association study for MacTel [132]. We could match eSNP:eGene pairs related to seven of the MacTel risk loci (chromosomes 1,2,5,7,10,11 and 17). Genes associated with the remaining loci were not expressed in either retina dataset. The number of matched eGene:eSNP pairs were summed for each of the four quadrants, representing the direction of effects, (Figure 3.8) Q1=+ve/+ve, Q2=+ve/-ve, Q3=-ve/+ve, Q4=-ve/-ve. Since the direction of the effects for the matched eQTLs were categorised into four quadrants, the cumulative counts in each quadrant for these seven loci were tested for concordance by estimating the unweighted Cohen's-kappa statistic (ranges from -1 to 1). This statistical method is mostly used to test interrater reliability and is not an inferential statistical test. Hence, using this method, we compared the agreement between the directions of effect for the matching MacTel loci using the results of the independent retinal eQTL studies i.e., Pinelli *et al.* with Ratnapriya *et al.* and Strunz *et al.* We observed poor concordance for all the comparisons made (Table 3.4), that is less than the conventional kappa > 0.7 for a satisfactory agreement [150]. There appears a better agreement while comparing Strunz *et al.* with Pinelli *et al.* and Ratnapriya *et al.* individually. However, there were relatively fewer matched counts compared to that of Pinelli *et al.* with Ratnapriya *et al.* and less reliable. This analysis indicates a lack of reproducibility for eQTL effects when comparing independent studies.

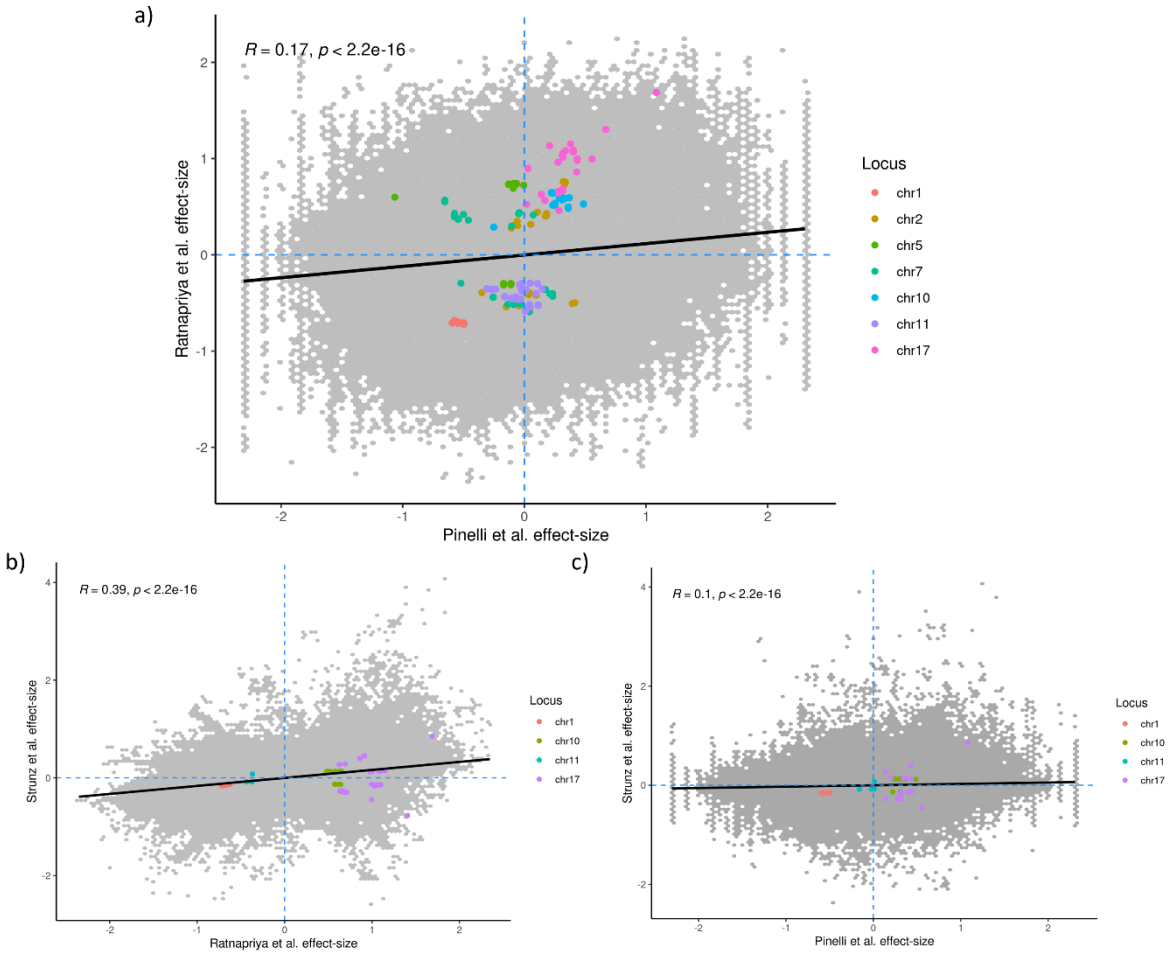

Figure 3.8: Comparing the concordance of the eQTL slopes identified for the MacTel associations using the EyeGEx (Ratnapriya *et al.*) with the matched eGene:eSNP pairs of the nominal results between a) Pinelli *et al.* vs. Ratnapriya *et al.* b) Ratnapriya *et al.* vs. Strunz *et al.* c) Pinelli *et al.* vs. Strunz *et al.* The blue lines demarcate the directions of paired effects (points on the plot) in each quadrant for the matched loci. The coloured points are slopes of the matched eQTLs at the MacTel loci and the grey blocks represent the background of all matched eQTL results.

Table 3.4: eQTL concordance for the direction of effect-sizes between three retinal eQTL studies.

Dataset compared	**Q1 (+/+)**	**Q2 (+/-)**	**Q3 (-/+)**	**Q4 (-/-)**	**Gene count in loci (eQTLs)**	**counts agreement proportion (1-alpha)**	**Cohen's Kappa**	**Lower CI**	**Higher CI**
Ratnapriya *et al.* with Pinelli *et al.*	118	54	101	114	18 (387)	0.599	0.211	0.116	0.304
Strunz *et al.* with Pinelli *et al.*	69	0	23	34	5 (126)	0.818	0.618	0.487	0.748
Ratnapriya *et al.* with Strunz *et al.*	187	54	124	148	5 (237)	0.852	0.678	0.583	0.774

Q: Quadrant; CI= Confidence Interval

3.4 Summary

We performed a retinal *cis*-eQTL analysis with a small but novel independent cohort. We identified four significant eQTLs at an FDR threshold of 0.1 following the procedure implemented by the GTEx consortium [91]. Two of the four significant eGenes, *AC091729.3* (7p22.3) and *LINC01535* (19q13.12) are long non-coding RNAs. *LINC01535* (19q13.12) is found to be involved in the progression of cervical cancer [152]. AC091729.3 has, as yet not been found to have any association with disease. The remaining two significant protein-coding eGenes were *HSD17B12* (11p11.2), *WFDC10B* (20q13.12). *HSD17B12* is involved in the fatty acyl-CoA biosynthesis pathway [153] and very long-chain fatty acid synthesis pathway [154] and an important gene to explore further since fatty acids are a source of energy in the retina. *WFDC10B* (20q13.12), which has been previously associated with AMD through a genome-wide association study [115] and is likely to progress AMD by its serine-type endopeptidase inhibitor activity digesting the extracellular matrix similar to *HTRA1*.

We compared our results with the previous retinal eQTL studies, Ratnapriya *et al.* and Strunz *et al.* The sample size of the Pinelli *et al.* study is much smaller (~10 %) than both studies. Therefore, we performed a statistical power estimation for eQTL data using the state-of-the-art eQTL power calculator to gain insights on the trade-off for the sample size to conventional power in the light of eQTLs. This analysis suggested that our data is underpowered for eQTL analysis and provided evidence of power estimation with real data for the theoretical study based on simulations for the eQTL power estimations [89].

We systematically modulated the eQTL input with a rationale to reduce the number of statistical tests to mitigate the multiple testing burden. However, whereas the relationship between the corrected P-values shifted in the expected direction, we have not observed an improvement in the eQTL signal for these runs. A restricted gene set analysis with subsets of genes identified through expression-based ranking for likely enhanced signal was performed. Further improvement in the statistical significance was not observed plausibly due to the decreased expression variance for highly expressed genes. Furthermore, exploring the results with a gene set of significantly associated genes from the GWASes of AMD & MacTel resulted in a slightly improved significance for genes associated with AMD.

From a technical standpoint, to address the underpowered cohort size, we tried several approaches to limit the multiple testing burden, namely applying higher MAF thresholds, restricting the number of phenotypes tested and performing analysis for genes ranked by their level of expression. While performing these analyses, we identified several areas that could be explored in future.

- The cis-eQTL identification window (± 1Mb of TSS) may not be adequate to identify the eQTLs for longer genes.
- An increasing MAF threshold implies reduced effect sizes of SNPs and so the P-values according to the complex trait genetics. Hence this might have unchanged the statistical significance as anticipated.
- Power analyses including the sample sizes from Ratnapriya *et al.* and Strunz *et al.* indicate an overperformance compared to the expected results, Ratnapriya *et al.* in particular, since the Strunz *et al.* despite incorporating the disease-free samples from Ratnapriya *et al.* achieved fewer results.

- A concordance analysis based on the direction of eQTL effects by independent studies showed less agreement inducing further exploration on effect-size reliability.

A recent study by Schwarz *et al.* [155] showed that large sample sizes are more powerful than deeper sequencing with a smaller cohort. Given retina is a rare tissue, the largest samples will be attainable through meta-analysis. Strunz *et al.* have performed similar analysis, however, further meta-analysis by combining Ratnapriya *et al.*, Strunz *et al.* with Orozco *et al.* [156] and our data will be worthwhile to expand our understanding of the genetic regulation in the retina.

Chapter 4: Regulation of retinal gene expression in MacTel

4.1 Introduction

The latest published GWA-study [132] for MacTel, published early in 2021 made significant contributions to the understanding of the aetiology of this retinal disorder. I contributed to this analysis, which is detailed in this chapter. This GWAS was performed with a dataset of 1,067 MacTel patients and 3,799 controls and confirmed the three previously reported risk loci in the transmembrane protein 161B (*TMEM161B*; 5q14.3), phosphoglycerate dehydrogenase (*PHGDH*; locus 1p12) and carbamoyl-phosphate synthase 1 (*CPS1*; 2q34) genes. It further revealed eight novel loci implicating processes such as serine-glycine metabolism, metabolite transport, and retinal vasculature and thickness, as well as five novel loci that attained suggestive significance (P < 5E-6). The total SNPs tested in the study and the eleven independent disease-associated SNPs that reached GW-significance representing the ten MacTel risk loci identified are presented in Figure 4.1. The details of the top-tagging SNPs for each of the MacTel disease-associated locus are reproduced from the publication (Table 4.1) for reference [132].

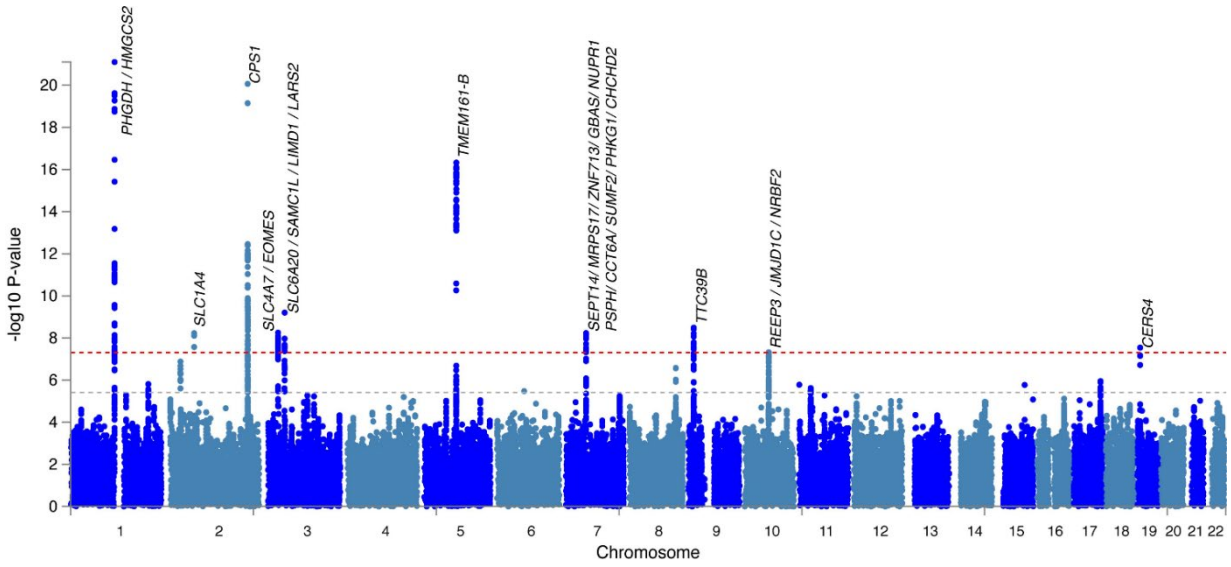

Figure 4.1: Manhattan plot displaying loci associated with macular telangiectasia type II. Each point denotes a single-nucleotide polymorphism (SNP) located on a particular chromosome (x-axis). The significance level (-log10 P-value) is represented on the y-axis. The dotted orange line indicates the threshold for genome-wide significance 5E-8 while the dotted grey line indicates the threshold for suggestive genome-wide significance 5E-6. The labels displayed for each locus are the genes located proximal to or within genome-wide significant loci. This is also Figure 2 from Bonelli, Jackson *et al.* [132].

Table 4.1: Top tagging GW-significant SNPs for each locus and relevant candidate genes from GWAS analysis. Bold text indicates the gene is covered by the haplotype (as defined by FUMA, based on linkage disequilibrium using the 1000 Genomes Project European cohort). The non-bold text indicates genes proximal to the locus. The Bonelli *et al.* GWAS [157] was analysed by conditioning on genetically predicted T2D risk, serum glycine and serine. Chr chromosome number, BP base pairs (Hg19 build), EAF effect allele frequency, CI confidence interval, GW genome-wide significant (P < 5E-8), S suggestive significant (P < 5E-6).

Rsid	**Chr**	**BP**	**effect/non-effect allele**	**EAF Cases**	**EAF Controls**	**Cytoband**	**Odds Ratio**	**95% C.I.**	**P-value**	**Proximal coding genes**	**Scerri *et al.* [127], 2017 GWAS**	**Bonelli *et al.* [157], conditional GWAS**
rs146953046	1	120278072	G/T	0.059	0.016	p12	5.47	3.87 - 7.74	7.90E-22	***PHGDH; HMGCS2***	-	-
rs532303	1	120265444	G/A	0.576	0.685	p12	0.585	0.52 - 0.65	9.29E-18	***PHGDH; HMGCS2***	GW	-
rs2160387	2	65220910	C/T	0.351	0.441	p14	0.72	0.65 - 0.81	5.90E-09	***SLC1A4***	-	-
rs1047891	2	211540507	A/C	0.192	0.306	q34	0.56	0.5 - 0.63	8.60E-21	***CPS1***	GW	-
rs9820465	3	27706298	C/T	0.145	0.205	p24.1	0.66	0.57 - 0.76	5.60E-09	*SLC4A7; EOMES*	S	GW
rs17279437	3	45814094	A/G	0.144	0.104	p21.31	1.73	1.45 - 2.06	6.20E-10	***SLC6A20; SACM1L; LIMD1; LARS2***	-	-
rs17421627	5	87847586	G/T	0.132	0.069	q14.3	2.31	1.9 - 2.81	4.70E-17	*TMEM161B*	GW	GW
rs6955423	7	56099352	A/G	0.651	0.746	p11.2	0.7	0.62 - 0.79	5.90E-09	***SEPT14;MRPS17;ZNF713;GBAS;PSPH***	S	-

										;CCT6A;SUMF2;PHKG1;CHCHD2;NUPR1L		
rs677622	9	15302613	G/A	0.821	0.869	p22.3	0.63	0.54 - 0.73	3.20E-09	***TTC39B***	S	-
rs10995566	10	65363166	T/C	0.231	0.316	q21.3	0.72	0.64 - 0.81	4.80E-08	***NRBF2;JMJD1C;REEP3***	S	-
rs139412173	19	8235251	G/A	0.024	0.056	p13.2	0.47	0.36 - 0.61	2.90E-08	*CERS4, FBN3*	S Nearby	GW

Further functional analysis to elucidate disease mechanisms underlying these GWAS signals is crucial to gain further biological insights. I contributed to this aspect of the study by performing colocalisation analysis of retinal eQTLs and MacTel risk loci using the tool *coloc* [99]. This tool can identify if the same variant is responsible for both transcriptional and disease phenotypes by comparing either GWAS-eQTL or GWAS-GWAS summary statistics [99]. The tool makes use of LD patterning of p-values which will be shared if the GWAS signals for both traits have arisen from the same causal variant. This variant may not be an eQTL, or an SNV but which is in LD with these two signals. Here, the GWA signal for the MacTel GW-significant loci that overlapped with 1) retinal eQTL data and 2) association results from other GWASes, were tested for colocalisation (details in methods).

In addition to colocalization analysis, bespoke eQTL analysis was required at the most significantly associated MacTel GWAS locus (encompassing the gene *PHGDH*). The strongest signal at this locus was for SNP rs146953046 (MAF = 1.6% in controls; 5.9% in cases, OR = 5.47). Since this SNP was found to be a significant eQTL for *PHGDH* in 27 tissues in the GTEx database (excluding retina), we analysed the possible eQTL effect of this relatively rare *PHGDH* SNP in the retina. To do so, we used retinal genotypes and RNA abundance data from Ratnapriya *et al.* [94]. The MacTel associated rare SNP was not included in the original publication of Ratnapriya *et al.*, necessitating the re-imputation of the genotype data and performing a very targeted eQTL analysis which I carried out as part of my thesis work.

4.2 Colocalization analysis using *coloc*

4.2.1 Methods

The SNPs for all the significant MacTel GWAS [132] risk loci were investigated via the FUMA portal [62] for possible associations with the expression of proximal genes (cis-eQTLs). Recently published retinal eQTLs from the EyeGEx resource (not available in FUMA) were also included after manual curation, together with specific eQTL datasets for the brain (16 regions), arteries (3 sites), tibial nerve, liver, EBV-transformed lymphocytes, and four whole-blood datasets (including GTEx, eQTLGen (eqtlgen.org) and BIOS consortia

data). At first, the MacTel effect alleles were reordered such that the disease effects were positive (i.e., increasing MacTel risk). Next, the effect alleles as well as the effect estimates were inverted in the retinal eQTL results to be consistent with the MacTel allele order and identified those genes for which the top SNP or any SNP in LD ($r^2 > 0.4$) was a significant eQTL, based on the gene-specific significance thresholds defined by Ratnapriya *et al.* [94]. The retinal eQTL signals at each GW-significant locus were analysed for colocalization using the R package *coloc* [99]. The effect sizes and corresponding variances from the GWAS and the eQTL analysis were given as input to the Approximate Bayes Factor function employed by *coloc* to estimate the posterior probability (PP) of a significant shared causal variant from GWAS and eQTL results. A PP > 0.75 was taken as the strong support for colocalization as recommended based on exploratory analysis from the study authors [99]. A colocalization test using this methodology was also performed for other published GWAS traits such as retinal vascular calibre (venular [158] and arteriolar [159]) and for serum Glycine and Serine abundance [160].

4.2.3 Results

Altogether we identified 99 genes in one or more tissues, whose expression was significantly affected by MacTel-associated SNPs. The most statistically significant SNP for MacTel was at the *PHGDH* locus (1p12). The matrix protein gene *MXRA7* (17q25.1) in six tissues, was the strongest result with a positive mean effect-size of 1.82. Other genes with statistically significant results were the zinc-finger domain protein *ZNF713* (7p11.2) and the mitochondrial respiration-related gene *GBAS* (7p11.2). Suppressive (negative) eQTL effects were inferred in several tissues of MacTel patients for the sphingolipid-related gene *SUMF2* (7p11.2) and differentiation factor *ATRAID* (2p23.3) genes. Importantly, suppression of *PHGDH* (1p12) expression was inferred in the brain, the tibial nerve and the vasculature. Interestingly, the direction of the eQTL transcriptional effect was opposite (positive in the brain, negative in the retina) for the genes *NRBF2* (10q21.3) and *TMEM161B* (5q14.3), and *DCDC1* (11p14.1).

The 2p14, 5q14.3, 7p11.2 and 10q21.3 GW-significant loci for MacTel overlapped with significant retinal eQTLs and were tested for colocalisation. The remaining GW-significant loci for MacTel 1p12, 2q34, 5q14.3, and 9p22.3 were tested for colocalisation

with the results from GWAS of MacTel-relevant traits. However, no evidence for shared underlying causal variants was found (i.e., all PP < 0.75). An example of a non-significant colocalization result between trait MacTel and serum glycine abundance at the *PHGDH* locus visualised using the R package LocusCompare [161], is provided below. It clearly shows that the top significant SNP (coloured purple) is different between the two GWAS-phenotypes (Figure 4.2). The plot also displays the haplotype structure for the locus from each study, with other SNPs that are in close LD being coloured by the strength of the LD relationship. Conversely, a perfect colocalization between traits would appear as an identical LD structure (right panels) and an x=y relationship between the two traits for the SNP P-values (left panel). The colocalization results of the four MacTel loci as well as the published GWAS traits are presented in the table below (Table 4.2).

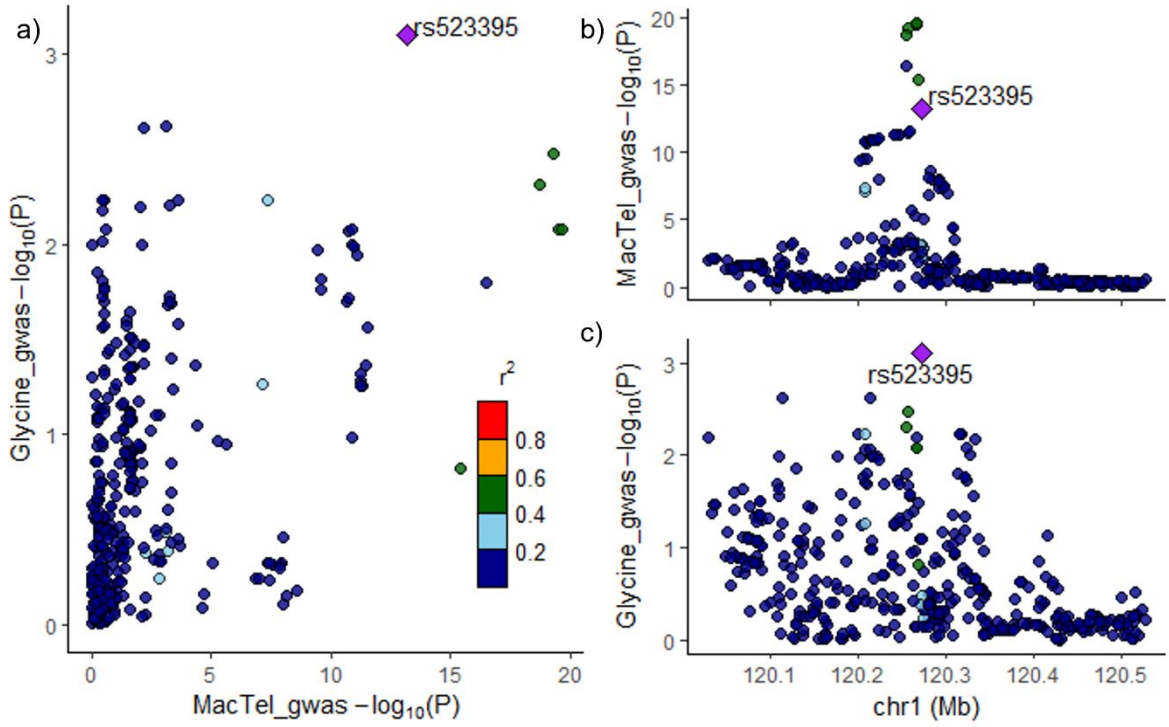

Figure 4.2: Example colocalisation plot that compares the GWAS results of the MacTel GWAS for the *PHGDH* locus (Chr1) with that of the serum glycine abundance from Shin *et al.* generated. a) Comparison GWAS p-values between the two studies with the points representing the SNPs coloured by the LD between them. Locus-zoom of the GWAS results from MacTel GWAS (b) and the GWAS results for serum glycine abundance (c) from Shin *et al.* with -log(P-value) on the y-axis and the locus of chromosome 1(Mb).

Table 4.2: Results of the colocalisation analysis performed for the significant MacTel loci using *coloc.*

					candidate gene			Co-localisation result		Top MacTel SNP intersecting trait SNP / eQTL				Top trait SNP / eQTL intersecting MacTel SNP			
Locus	**Study type**	**Trait**	**Reference publication**	**Colocalized SNP**	**ENSEMBL ID**	**symbol**	**N. SNPs evalua-ted**	**PP shared causal variants**	**PP distinct causal variants**	**rsID**	**CHR :POS**	**MacTel GWAS p-val**	**Retina eQTL p-val ***	**rsID**	**CHR: POS**	**MacTel GWAS p-val**	**Retina eQTL p-val ***
2p14	eQTL	Retinal gene expression	Ratnapriya *et al.* 2019 Nat Genet	rs2422358	ENSG00000115902	SLC1A4	1199	0.0868	99.2	rs2422358	2:65231806	7.84E-09	3.35E-07	rs7566124	2:65224470	9.34E-02	4.74E-26
5q14.3				rs12517664	ENSG00000164180	TMEM161B	667	3.75	52.4	rs17480689	5:87834883	7.58E-17	1.81E-01	rs11748762	5:87650585	3.68E-01	2.36E-06
				rs16903178	ENSG00000247828	TMEM161B-AS1	667	0.000000747	100	rs17480689	5:87834883	7.58E-17	4.45E-03	rs112498091	5:87645835	1.14E-04	5.21E-24
7p11.2				rs12669623	ENSG00000129103	SUMF2	1534	0.0058	99.9	rs6955423	7:56099352	5.88E-09	5.21E-05	rs4427122	7:56117707	4.84E-01	7.51E-22
				rs12669623	ENSG00000178665	ZNF713	1534	0.0101	99.9	rs6955423	7:56099352	5.88E-09	2.69E-05	rs66741911	7:56050170	3.33E-01	6.67E-18
				rs10081373	ENSG00000146729	GBAS	1534	0.164	99.8	rs6955423	7:56099352	5.88E-09	1.74E-01	rs62457265	7:56111710	1.06E-03	5.98E-19
10q21.3				rs10995566	ENSG00000148572	NRBF2	2303	6.81	92.7	rs10995566	10:65363166	4.78E-08	3.56E-16	rs10995477	10:65010672	2.77E-04	3.08E-27

				rs7895549	ENSG00000171988	JMJD1C	2303	4.56	70.8	rs10995566	10:65363166	4.78E-08	2.57E-01	rs10740138	10:65395133	4.57E-04	4.13E-05
1p12	GWAS	Serum serine abundance	Shin *et al.* 2014 Nat Genet	rs532303	ENSG00000092621	PHGDH	417	40.6	59.4	rs532303	1:12026690 3	2.36E-20	7.82E-24	rs1163251	1:12020975 5	2.24E-11	7.05E-27
2q34				rs715	ENSG00000021826	CPS1	515	99	1	rs715	2:21154305 5	7.11E-20	2.70E-21	rs715	2:21154305 5	7.11E-20	2.69E-21
		Serum glycine abundance		rs715	ENSG00000021826	CPS1	513	100	0	rs715	2:21154305 5	7.11E-20	7.10E-20	rs715	2:21154305 5	7.11E-20	7.11E-20
5q14.3		Retinal Calibre (venular)	Ikram *et al.* 2010 PLoS Genet	rs17421627	ENSG00000245526	LINC00461	770	62.5	6	rs17421627	5:87847586	4.70E-17	5.05E-09	rs17421627	5:87847586	4.70E-17	5.05E-09
		Retinal Calibre (arteriolar)	Sim *et al.* 2013 PloS One	rs2194025	ENSG00000271904	AC091826.3	753	94	4.8	rs17421627	5:87847586	4.70E-17	6.46E-10	rs2194025	5:87798236	9.07E-15	1.53E-10
7p11.2		Serum serine abundance	Shin *et al.* 2014 Nat Genet	rs4689	ENSG00000146729	NIPSNAP2	255	99	1	rs13241866	7:56090878	1.08E-08	2.28E-04	rs10229446	7:56064262	3.28E-08	1.52E-13
PP shared causal variants - Posterior probability that the Mactel GWAS and Retina eQTL signals derive from a shared variant.																	
PP distinct causal variants - Posterior probability that the Mactel GWAS and Retina eQTL signals derive from distinct variants.																	
* P-values in bold meet the gene-specific threshold for a significant eQTL as determined by Ratnapriya *et al.* Nature Genetics 2019.																	

4.3 eQTL analysis for the rare *PHGDH* MacTel-associated SNP

4.3.1 Methods

To investigate potential eQTL effects for the rare *PHGDH* SNP rs146953046, which was not reported in the original retinal eQTL study by Ratnapriya *et al.* [94] (cohort size of 406 individuals), the genotype data for that study was generously provided by Dr. Rinki Ratnapriya and Prof. Anand Swaroop (US National Eye Institute). Updated genotypes were imputed using the Sanger Imputation Server version (EAGLE2 v2.0.5 + PBWT) [138], [162] and the Haplotype Reference Consortium panel as a reference (r1.1). The corresponding retinal RNA-seq data (GEO accession GSE115828) was aligned to the human genome (release GRCh38.91) using the STAR alignment tool (ver. 2.6.1) [41], and transcriptional abundance was calculated using featureCounts (ver.2.0.0) [43]. Linear regression controlling for age, sex, and AMD disease status were employed to examine the expression of *PHGDH* in individuals heterozygous for rs146953046 relative to homozygous reference individuals.

To test the potential splicing effect of this SNP, we searched the GTEx portal. Next, differential exon usage analysis was performed on the Ratnapriya *et al.* data by merging overlapping exons and quantifying their abundance using the SubRead module's flattenGTF and featureCounts respectively [43]. The exons with fewer than 1 count per million in at least 40 subjects were discarded. Differential exon abundance was tested using the limma diffSplice module [163]. Fold-change and nominal P values for exons of *PHGDH* were extracted for graphical analysis.

4.3.2 Results

The possible impact on retinal transcription of the rare and deleterious SNP rs146953046 at the *PHGDH* locus was investigated via the GTEx server, where this SNP was a significant eQTL for *PHGDH* in 27 tissues (including 12 brain regions). We found that all normalised effects of the effect (G) allele suppressed gene expression (Figure 4.3a). As this SNP was not included in the original retinal eQTL study by Ratnapriya *et al.* [94], the genotype data from that study was re-imputed (methods), from which we identified two healthy controls and seven age-related macular degeneration (AMD) patients as

heterozygotes for this SNP. No homozygotes for the alternate allele were identified. Significantly lower *PHGDH* expression was observed in the subjects heterozygous for this SNP, compared to homozygous reference subjects when correcting for age, sex and AMD status ($P < 0.003$; Figure 4.3b). Given that spliceQTL effects are also reported for rs146953046 in the skin, tibial nerve and artery, and oesophagal mucosa, we compared *PHGDH* exon expression in the retina between heterozygotes and reference homozygous subjects using a similar statistical framework as for differential gene expression testing but, found no significant differences (Figure 4.2c).

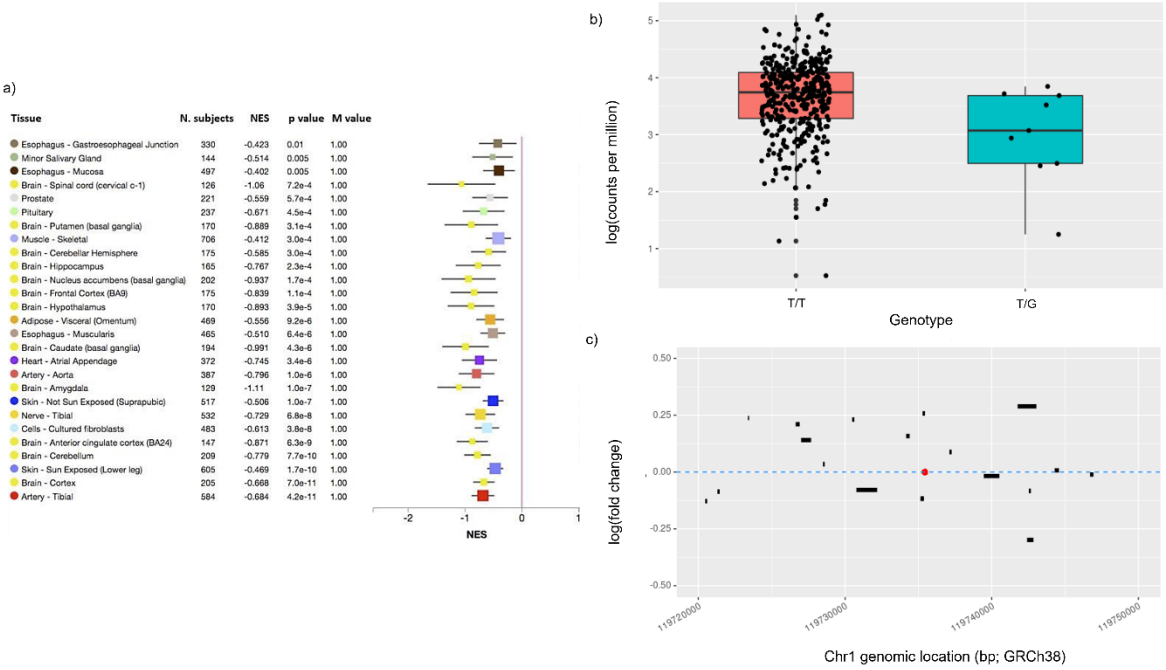

Figure 4.3: Effects of the rare deleterious SNP rs146953046 on gene expression and exon abundance in the GTEx database and the retina. a) Tissues in the GTEx database (y-axis) for which the minor (G) allele of rs146953046 is associated with a significant difference in *PHGDH* gene expression with high confidence (M value =1). NES: normalized effect size (equivalent to alternative SNP beta value in a linear model); M value: the posterior probability of eQTL. Horizontal whisker plots indicate the median (filled square) and 95% confidence interval (black bars) for NES estimates. In all tissues the effect of the G allele is suppressive. b) Expression of *PHGDH* in the retina (y-axis) for individuals with the reference genotype (T/T) compared to heterozygotes (x-axis). Each point represents gene expression of *PHGDH* measured in an individual donor

retina. Box plot centre line: median; box limits: upper and lower quartiles; whiskers: 1.5x interquartile range. The difference between genotypes is significant ($p < 0.003$) after correcting for subject age, sex, and AMD status. c) Relative expression of merged *PHGDH* exons in the neural retina between heterozygous individuals and reference homozygous individuals. The span of each exon across chromosome 1 (x-axis) is indicated by black bars. The height of the exon bars relative to the y axis indicates log fold change. The red point indicates the genomic (intronic) location of the rare SNP.

4.4 Summary

In this chapter, I presented the analysis performed related to the GWAS for MacTel, a relatively rare retinal disease with a cohort consisting of 1067 MacTel patients and 3799 controls [132]. We identified 11 genome-wide associated SNPs at 10 loci replicating the three loci previously identified in the first GWAS of MacTel. The serine and glycine pathway dysregulation has been further pinpointed to be a putative cause of MacTel as eluded in the first MacTel GWAS [127]. The most significant disease signals observed in this MacTel GWAS study are located within the *PHGDH* gene locus (1p12). The encoded enzyme phosphoglycerate dehydrogenase is a crucial rate-limiting enzyme in the serine biosynthesis pathway. The significant GW-associated variants indicate the reduced activity of *PHGDH* and thereby affecting the serine synthesis crucial for the retina. The reduced levels of circulating serine in the serum, as demonstrated in the mice retina [133], resulted in the accumulation of the neurotoxins (deoxy-sphingolipids) in the retina, whose depositions correlated with the MacTel phenotype [133]. The retinal pigmented epithelium (RPE) located at the interface between the photoreceptors and the choroid capillaries supports the neural retina by transporting nutrients including serine. RPE cells are enriched for sphingolipids and are crucial in the survival of photoreceptors by recycling their membrane lipids (chapter 1). The MacTel-variants associated with the *PHGDH* gene have been found to elevate the levels of deoxySphingolipids in the RPE [133], [164]–[166]. The rare risk haplotype, tagged by the SNP rs146953046, is associated with lower *PHGDH* expression in the retina. The common GW-significant SNP (tagged by rs532303) was not identified as a significant regulatory variant (eQTL) in the retina [94]. Although no evidence of colocalization was observed for

the four GW-significant MacTel loci tested, the identification of two independent SNPs associated with *PHGDH*, and support for the involvement of this gene in the MacTel study results from SNP enrichment, eQTL and TWAS results [132], indicates a central but complex role in disease manifestation. Another study on MacTel identified 22 rare variants in the *PHGDH* gene in the MacTel subjects and determined that the genetic architecture of MacTel is highly heterogeneous with no other genes except PHGDH explaining more than 0.5% of the disease [167].

With regards to the investigation of splicing effects of the *PHGDH* rare SNP rs146953046, there are four tissues, the skin, tibial nerve and artery, and oesophagal mucosa, in GTEx for which this SNP is both an eQTL and a splicing QTL. The differential exon usage analysis provided no evidence of the splicing effect in the retina for this SNP. However, in chapter 2 we observed that the Ratnapriya *et al.* retinal RNA-seq data contains relatively small read library sizes which may reduce the power of this analysis. We explored Pinelli *et al.* for the rare *PHGDH* SNP, however, only one of the samples is observed to be a heterozygote and provided little evidence to examine further. Despite this, these results enhance the likelihood of the SNP rs146953046 to be disease-causing with a provisional mechanism of reduced *PHGDH* gene expression, thereby suppressing serine biosynthesis in the retina.

A recent study using mouse models with serine depleted diet could sufficiently generate a MacTel-like phenotype and in conjunction with the findings from our study suggest that adequate serum serine abundance is required to maintain retinal health [133]. Although MacTel is a rare disease, this GWAS study assists the research community to improve the understanding of MacTel aetiology as well as informing about important metabolic and tissue-specific processes of Retina as well as the adjacent cells like RPE.

Chapter 5: Discussion

The retina is the light-sensitive ocular tissue responsible for vision. Studying the genetic basis of the debilitating heritable retinal disorders, such as Age-related macular degeneration (AMD), and Macular telangiectasia II (MacTel) is important to gain functional genomic insights into the aetiology of these diseases affecting the Macula, the central-most region of the retina responsible for high acuity vision. Identifying essential regulators of gene expression in the healthy and diseased retina continues to be a significant challenge in macular degeneration research. Genome-wide genetic association studies help us to identify the genetic risk loci associated with such complex genetic diseases where many variants contribute risk to the disease. However, little is known about the genetic regulation of retinal gene expression. Functional studies such as eQTLs, elucidate the relationship of genetic changes and their influence on the expression of genes. eQTLs are widely used to interpret GWAS signals, to identify driver genes and biological mechanisms. With the goal of improving our understanding of the regulation of gene expression in various tissues, the GTEx [68], [91] project provides comprehensive information about eQTLs, sQTLs and transcriptomes of various human tissues, however, the retina was missing in the set of >50 tissues that comprised the original GTEx tissues. Due to the difficulties of obtaining human retinal tissue, only a few studies, such as Ratnapriya *et al.*, 2019 [94] and Strunz *et al.*, 2020 [136] using AMD affected and disease-free individuals respectively, have analysed the functional genomic aspects of the retina. The first retinal transcriptome was assembled by Pinelli *et al.* [135] in 2016 and the associated genotype data was generously provided by the authors to conduct the work described in this thesis. The main goal of this thesis is to explore the relationship of genetic change with gene expression in the retina and to interpret the biology of genome-wide association results for retinal diseases such as AMD and MacTel. Thus, we focused on the *cis*-eQTL identification with the cohort of disease-free retinae from Pinelli *et al.* as an independent validation of the recent studies. Further, we compared our results with the published studies.

As detailed in chapter 2, the transcriptomic data and SNP genotype data for the disease-free individuals were integrated using linear regression to perform eQTL analysis in

the retina using QTLtools. We found that our cohort is very homogenous with European (EUR) ethnicity without any population stratification by comparing against the genotypes from the Thousand Genome project [17]. Since the prevalence of retinal diseases varies with ethnicity, this is a useful property of our data to study EUR population compared to the published studies which used a mixed population. Other advantages of this data were the large number of high-quality SNPs retained after quality control and post-imputation quality assessment, and the read depth of gene expression data, which was twice as deep as either the Strunz or Ratnapriya *et al.* datasets. The major disadvantage was a modest sample size (n=41) compared to the published studies and this diminished the advantages of this dataset. However, the retinal eQTLs obtained, after following the workflow presented in chapter 3, were modest and at an FDR < 0.1, we only found four statistically significant eGenes. The large discrepancy of significant results, in comparison to other studies, could be attributed to the huge difference in sample size and its influence on reaching adequate statistical power. Curiously the Strunz *et al.* study, with nearly three-quarters of the sample size and included the disease-free samples from the Ratnapriya *et al.* study, had achieved just over a quarter of the significant eGenes (compared to Ratnapriya *et al.* despite following the same analysis workflow except for the difference in eQTL mapping tool i.e., FastQTL [88] by Strunz *et al.* and QTLtools [77] by Ratnapriya *et al.*) demonstrates that sample size is not the only important factor in determining the number of significant eQTLs. This inconsistent relationship between sample size and eQTL discoveries points to the strong impact of other factors such as the differences in analysis workflows or differences in the statistical thresholds being employed (Table 3.2).

Two out of the four significant eGenes were *AC091729.9* (7p22.3), an uncharacterised long non-coding RNA gene and *LINC01535* (19q13.12), a long non-coding RNA gene found to be associated with cervical cancer [152]. From the perspective of MacTel, a significant GWAS SNP identified in the locus 5q14.3, is a highly conserved variant, modulating the expression of microRNA miR-9-5 in zebrafish [60]. Depletion of this locus in zebrafish resulted in retinal vasculature defects. Such a link with non-coding elements in the retina strongly pinpoints the vascular defects attributed to MacTel. While the other two were protein-coding genes. *HSD17B12* (11p11.2), a hydroxy-steroid(beta 12)

dehydrogenase 12, was found to be involved in the fatty acyl-CoA biosynthesis pathway [153] and a focal point for the very-long-chain fatty acid synthesis pathway [154]. Further exploring the associations of this gene is of particular interest since fatty acids are a preferred source of energy for the retina through oxidation, apart from glucose, and dysregulated lipid metabolic pathways affect the retinal health presenting diseases like AMD and MacTel [120], [133], [134]. *WFDC10B* (20q13.12), a WAP four-disulfide core domain protein, is a candidate gene associated with AMD [115]. *WFDC10B* has a Gene Ontology (GO) of serine-type endopeptidase inhibitor activity and tentatively progress AMD through excessive fragmenting of the retinal extracellular matrix in a likely mechanism, as presented by May *et al.* [168], of *HTRA1* mediated AMD pathogenesis. Although only four significant eGenes suggested a weak eQTL signal from the cohort data, all these four genes were found to be significant eGenes in at least one of the studies, Ratnapriya *et al.* or Strunz *et al.*, that published retinal eQTLs.

Given the four significant results, we further explored the data from Pinelli *et al.* 1) Can we detect a better signal with a sample size of 41 individuals? 2) Can we improve the signal by reducing the number of tests, limiting either the number of eSNPs by MAF, or genes by expression and disease association? To answer the first question, we performed an extensive power study as detailed in chapter 3. The issue of lower statistical power for our cohort is further confirmed by an extensive simulation study on eQTLs previously performed [89]. As part of this thesis, we have explored the state-of-the-art eQTL power calculator, powerEQTL [151], which emphasised the need for a greater sample size. This implicates the primacy of sample size over the depth of the gene-expression abundance for obtaining power in eQTL studies. Power estimations in the light of sample size were not addressed in either of the previous retinal eQTL studies and our exploration of power indicates an overperformance relative to the expected retinal eQTL results, Ratnapriya *et al.*, in particular.

Subsequently, to explore the second question of improvement of eQTL signal, with anticipation of improved statistically significant results, analyses were performed with a systematic reduction of the genotype data applying higher MAF thresholds (such as 0.05, 0.1, 0.2, 0.3, 0.4), as a potential way of mitigating the multiple testing burden. Upon

comparing the results obtained for each MAF threshold, although we observed an improved P-value significance yet, no further improvement after FDR correction was observed.

To complement the previous approach, we also restricted the number of phenotypes (genes) being tested to attempt to improve statistical power. Of the seven statistically significant eGenes identified through this approach, all were candidate genes for AMD. Four out of these seven genes were found to be significant eGenes in either of the published retinal eQTL studies [94], [136]. This result indicates that the Pinelli *et al.* data contains the anticipated genetic signal but resulted in an overall weak statistical significance, mainly driven by the small sample size.

A recent largest GWAS study on MacTel identified ten genome-wide significant loci and six suggestively significant loci. Retinal eQTLs within these risk loci were identified using the published retinal eQTLs (Ratnapriya *et al.*, 2019 [94]). We were curious to observe the reliability of the effect-sizes for the eQTLs matched within these loci from the Pinelli *et al.* results, by analysing the direction of their effect. As detailed in Chapter 3, we found that the concordance between studies was very modest. This lack of concordance in the effect-sizes forecasts the difficulties of replicating the eQTL effects from independent studies. However, other studies have also noted that eQTL effects vary widely across various tissues [91], [169].

As discussed in chapter 4, we performed a targeted eQTL analysis to examine if the strongest GWAS signal from MacTel, *rs146953046*, a rare SNP of *PHGDH* gene, was an eQTL in the retina, since it was found to be an eQTL in 27 different tissues in GTEx [68], [91]. This necessitated the imputation of genotype data provided by Ratnpriya *et al.* [94] since this important SNP was missing in the original imputation dataset. We observed a suppressive association of the novel, rare, MacTel risk allele on the *PHGDH* gene. Further extension of these efforts to check for potential splicing effects (as *rs146953046* is also an sQTL in four tissues) by performing a differential exon usage analysis, found no significant splicing effect by this SNP, likely due to the smaller library size and rarity of the SNP. Our efforts to identify a co-localisation of the MacTel GWAS signals using *coloc [99]* not only

with the retinal eQTLs but also with other related quantitative traits such as retinal vascular calibre and serum metabolites found no evidence for colocalisation, implying a particular mechanism of the variant to MacTel as pinpointed by this MacTel GWAS. However, two independent SNPs associated with *PHGDH*, and its reduced activity resulting in the accumulation of neuro-toxins, thereby generating the MacTel phenotype, strengthens the evidence for its crucial role in MacTel disease aetiology [132], [133], [167].

Future work

This work focused on identifying eQTLs in a novel human retinal dataset. Our cohort represents one of the few independent datasets available to study retina, which is a rare resource. We have demonstrated the utility of such integrated studies to enhance the understanding of functional genomics aspects of common complex disorders. In future, this work could be expanded to interpret future GWAS or extrapolate to further investigate published GWASes. Although we attained modest statistical significance due to the limiting sample size, performing a meta-analysis by incorporating the Pinelli *et al.* data along with the published studies to further improve the power of retinal eQTLs discovery will be worthwhile. The robust read depth of our data and the fact that splicing QTLs are highly compatible amongst tissues can be further harnessed to explore splicing QTLs in retinal tissue. Since sQTLs are observed in the brain tissue [170] and the retinal transcriptome is closer to the brain [94], we expect that the high informational reads of our data yields more statistical power to identify sQTLs in the retina and our dataset will have advantages over published retina studies. Furthermore, sQTLs have thus far not yet been explored for the retina.

An important extension to all future eQTL studies is to move from bulk RNAseq to single-cell level to identify the various cell types that mediate genetic effects. Such datasets combined with the methods employed here will increase our understanding of the regulation of gene expression in the retina and its contribution to disease.

References

[1] H. Kolb, E. Fernandez, and R. Nelson, Eds., *Webvision: The Organization of the Retina and Visual System*. Salt Lake City (UT): University of Utah Health Sciences Center, 2011. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/21413389

[2] B. D. Kels, A. Grzybowski, and J. M. Grant-Kels, "Human ocular anatomy," *Clin. Dermatol.*, vol. 33, no. 2, pp. 140–146, Mar. 2015, doi: 10.1016/j.clindermatol.2014.10.006.

[3] J. Stein-Streilein, "Immune regulation and the eye," *Trends Immunol.*, vol. 29, no. 11, pp. 548–554, Nov. 2008, doi: 10.1016/j.it.2008.08.002.

[4] P. E. Ludwig, R. Jessu, and C. N. Czyz, "Physiology, Eye," in *StatPearls*, Treasure Island (FL): StatPearls Publishing, 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/29262001

[5] D. Purves *et al.*, *The Retina*. Sinauer Associates, 2001. Accessed: Jan. 16, 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK10885/

[6] T. G. Wensel, "Molecular Biology of Vision," in *Basic Neurochemistry (Eighth Edition)*, S. T. Brady, G. J. Siegel, R. W. Albers, and D. L. Price, Eds. New York: Academic Press, 2012, pp. 889–903. doi: 10.1016/B978-0-12-374947-5.00051-1.

[7] S. Kawamura and S. Tachibanaki, "Rod and cone photoreceptors: molecular basis of the difference in their physiology," *Comp. Biochem. Physiol. A Mol. Integr. Physiol.*, vol. 150, no. 4, pp. 369–377, Aug. 2008, doi: 10.1016/j.cbpa.2008.04.600.

[8] D.-G. Luo, T. Xue, and K.-W. Yau, "How vision begins: an odyssey," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 105, no. 29, pp. 9855–9862, Jul. 2008, doi: 10.1073/pnas.0708405105.

[9] T. G. Kusakabe, N. Takimoto, M. Jin, and M. Tsuda, "Evolution and the origin of the visual retinoid cycle in vertebrates," *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, vol. 364, no. 1531, pp. 2897–2910, Oct. 2009, doi: 10.1098/rstb.2009.0043.

[10] J. C. Saari and D. L. Bredberg, "Lecithin:retinol acyltransferase in retinal pigment epithelial microsomes," *J. Biol. Chem.*, vol. 264, no. 15, pp. 8636–8640, May 1989, [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/2722792

[11] A. Tsin, B. Betts-Obregon, and J. Grigsby, "Visual cycle proteins: Structure, function, and roles in human retinal disease," *J. Biol. Chem.*, vol. 293, no. 34, pp. 13016–13021, Aug. 2018, doi: 10.1074/jbc.AW118.003228.

[12] M. Zarbin, "Cell-Based Therapy for Degenerative Retinal Disease," *Trends Mol. Med.*, vol. 22, no. 2, pp. 115–134, Feb. 2016, doi: 10.1016/j.molmed.2015.12.007.

[13] B. R. Korf, "Chapter 16 - Introduction to Human Genetics∗," in *Clinical and Translational Science (Second Edition)*, D. Robertson and G. H. Williams, Eds. Academic Press, 2017, pp. 281–311. doi: 10.1016/B978-0-12-802101-9.00016-8.

[14] "National Human Genome Research Institute Home | NHGRI." https://www.genome.gov/ (accessed Jul. 17, 2020).

[15] "Human Genome Project: Sequencing the Human Genome | Learn Science at Scitable." https://www.nature.com/scitable/topicpage/dna-sequencing-technologies-key-to-the-human-828/ (accessed Jul. 17, 2020).

[16] International Human Genome Sequencing Consortium, "Finishing the euchromatic sequence of the human genome," *Nature*, vol. 431, no. 7011, pp. 931–945, Oct. 2004, doi: 10.1038/nature03001.

[17] 1000 Genomes Project Consortium *et al.*, "A map of human genome variation from population-scale sequencing," *Nature*, vol. 467, no. 7319, pp. 1061–1073, Oct. 2010, doi: 10.1038/nature09534.
[18] L. Bull, "Genetics, Mutations, and Polymorphisms," 2013, Accessed: Jun. 03, 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK6475/?report=printable
[19] K. J. Karczewski *et al.*, "The mutational constraint spectrum quantified from variation in 141,456 humans," *Nature*, vol. 581, no. 7809, pp. 434–443, May 2020, doi: 10.1038/s41586-020-2308-7.
[20] R. Sachidanandam *et al.*, "A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms," *Nature*, vol. 409, no. 6822, pp. 928–933, Feb. 2001, doi: 10.1038/35057149.
[21] A. Chandra, D. Mitry, A. Wright, H. Campbell, and D. G. Charteris, "Genome-wide association studies: applications and insights gained in Ophthalmology," *Eye* , vol. 28, no. 9, pp. 1066–1079, Sep. 2014, doi: 10.1038/eye.2014.145.
[22] P. M. Visscher *et al.*, "10 Years of GWAS Discovery: Biology, Function, and Translation," *Am. J. Hum. Genet.*, vol. 101, no. 1, pp. 5–22, Jul. 2017, doi: 10.1016/j.ajhg.2017.06.005.
[23] M. J. Landrum *et al.*, "ClinVar: improving access to variant interpretations and supporting evidence," *Nucleic Acids Res.*, vol. 46, no. D1, pp. D1062–D1067, Jan. 2018, doi: 10.1093/nar/gkx1153.
[24] L. Clarke *et al.*, "The 1000 Genomes Project: data management and community access," *Nat. Methods*, vol. 9, no. 5, pp. 459–462, Apr. 2012, doi: 10.1038/nmeth.1974.
[25] M. Slatkin, "Linkage disequilibrium--understanding the evolutionary past and mapping the medical future," *Nat. Rev. Genet.*, vol. 9, no. 6, pp. 477–485, Jun. 2008, doi: 10.1038/nrg2361.
[26] J. Perkel, "SNP genotyping: six technologies that keyed a revolution," *Nat. Methods*, vol. 5, no. 5, pp. 447–453, May 2008, doi: 10.1038/nmeth0508-447.
[27] S. Das, G. R. Abecasis, and B. L. Browning, "Genotype Imputation from Large Reference Panels," *Annu. Rev. Genomics Hum. Genet.*, vol. 19, pp. 73–96, Aug. 2018, doi: 10.1146/annurev-genom-083117-021602.
[28] Y. Li, C. Willer, S. Sanna, and G. Abecasis, "Genotype imputation," *Annu. Rev. Genomics Hum. Genet.*, vol. 10, pp. 387–406, 2009, doi: 10.1146/annurev.genom.9.081307.164242.
[29] Y. Li, C. J. Willer, J. Ding, P. Scheet, and G. R. Abecasis, "MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes," *Genet. Epidemiol.*, vol. 34, no. 8, pp. 816–834, Dec. 2010, doi: 10.1002/gepi.20533.
[30] S. R. Browning and B. L. Browning, "Haplotype phasing: existing methods and new developments," *Nat. Rev. Genet.*, vol. 12, no. 10, pp. 703–714, Sep. 2011, doi: 10.1038/nrg3054.
[31] International HapMap Consortium, "The International HapMap Project," *Nature*, vol. 426, no. 6968, pp. 789–796, Dec. 2003, doi: 10.1038/nature02168.
[32] International HapMap Consortium, "A haplotype map of the human genome," *Nature*, vol. 437, no. 7063, pp. 1299–1320, Oct. 2005, doi: 10.1038/nature04226.
[33] International HapMap Consortium *et al.*, "A second generation human haplotype map of over 3.1 million SNPs," *Nature*, vol. 449, no. 7164, pp. 851–861, Oct. 2007, doi:

10.1038/nature06258.
[34] International HapMap 3 Consortium *et al.*, "Integrating common and rare genetic variation in diverse human populations," *Nature*, vol. 467, no. 7311, pp. 52–58, Sep. 2010, doi: 10.1038/nature09298.
[35] UK10K Consortium *et al.*, "The UK10K project identifies rare variants in health and disease," *Nature*, vol. 526, no. 7571, pp. 82–90, Oct. 2015, doi: 10.1038/nature14962.
[36] S. McCarthy *et al.*, "A reference panel of 64,976 haplotypes for genotype imputation," *Nat. Genet.*, vol. 48, no. 10, pp. 1279–1283, Oct. 2016, doi: 10.1038/ng.3643.
[37] B. N. Howie, P. Donnelly, and J. Marchini, "A flexible and accurate genotype imputation method for the next generation of genome-wide association studies," *PLoS Genet.*, vol. 5, no. 6, p. e1000529, Jun. 2009, doi: 10.1371/journal.pgen.1000529.
[38] B. L. Browning, Y. Zhou, and S. R. Browning, "A One-Penny Imputed Genome from Next-Generation Reference Panels," *Am. J. Hum. Genet.*, vol. 103, no. 3, pp. 338–348, Sep. 2018, doi: 10.1016/j.ajhg.2018.07.015.
[39] S. Das *et al.*, "Next-generation genotype imputation service and methods," *Nat. Genet.*, vol. 48, no. 10, pp. 1284–1287, Oct. 2016, doi: 10.1038/ng.3656.
[40] Z. Wang, M. Gerstein, and M. Snyder, "RNA-Seq: a revolutionary tool for transcriptomics," *Nat. Rev. Genet.*, vol. 10, no. 1, pp. 57–63, Jan. 2009, doi: 10.1038/nrg2484.
[41] A. Dobin *et al.*, "STAR: ultrafast universal RNA-seq aligner," *Bioinformatics*, vol. 29, no. 1, pp. 15–21, Jan. 2013, doi: 10.1093/bioinformatics/bts635.
[42] S. Anders, P. T. Pyl, and W. Huber, "HTSeq--a Python framework to work with high-throughput sequencing data," *Bioinformatics*, vol. 31, no. 2, pp. 166–169, Jan. 2015, doi: 10.1093/bioinformatics/btu638.
[43] Y. Liao, G. K. Smyth, and W. Shi, "featureCounts: an efficient general purpose program for assigning sequence reads to genomic features," *Bioinformatics*, vol. 30, no. 7, pp. 923–930, Apr. 2014, doi: 10.1093/bioinformatics/btt656.
[44] M. G. Grabherr *et al.*, "Full-length transcriptome assembly from RNA-Seq data without a reference genome," *Nat. Biotechnol.*, vol. 29, no. 7, pp. 644–652, May 2011, doi: 10.1038/nbt.1883.
[45] C. W. Law, M. Alhamdoosh, S. Su, G. K. Smyth, and M. E. Ritchie, "RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR," *F1000Res.*, vol. 5, no. 1408, p. 1408, Nov. 2016, doi: 10.12688/f1000research.9005.2.
[46] S. B. Montgomery, T. Lappalainen, M. Gutierrez-Arcelus, and E. T. Dermitzakis, "Rare and common regulatory variation in population-scale sequenced human genomes," *PLoS Genet.*, vol. 7, no. 7, p. e1002144, Jul. 2011, doi: 10.1371/journal.pgen.1002144.
[47] K. R. Kukurba and S. B. Montgomery, "RNA Sequencing and Analysis," *Cold Spring Harb. Protoc.*, vol. 2015, no. 11, pp. 951–969, Apr. 2015, doi: 10.1101/pdb.top084970.
[48] M. F. Bennett *et al.*, "Familial adult myoclonic epilepsy type 1 SAMD12 TTTCA repeat expansion arose 17,000 years ago and is present in Sri Lankan and Indian families," *Eur. J. Hum. Genet.*, vol. 28, no. 7, pp. 973–978, Jul. 2020, doi: 10.1038/s41431-020-0606-z.
[49] H. Rafehi *et al.*, "Bioinformatics-Based Identification of Expanded Repeats: A Non-reference Intronic Pentamer Expansion in RFC1 Causes CANVAS," *Am. J. Hum.*

Genet., vol. 105, no. 1, pp. 151–165, Jul. 2019, doi: 10.1016/j.ajhg.2019.05.016.
[50] "Phenotype Variability: Penetrance and Expressivity." https://www.nature.com/scitable/topicpage/phenotype-variability-penetrance-and-expressivity-573/ (accessed Apr. 19, 2021).
[51] J. E. Bailey-Wilson and A. F. Wilson, "Linkage analysis in the next-generation sequencing era," *Hum. Hered.*, vol. 72, no. 4, pp. 228–236, Dec. 2011, doi: 10.1159/000334381.
[52] M. I. McCarthy *et al.*, "Genome-wide association studies for complex traits: consensus, uncertainty and challenges," *Nat. Rev. Genet.*, vol. 9, no. 5, pp. 356–369, May 2008, doi: 10.1038/nrg2344.
[53] K. T. Zondervan and L. R. Cardon, "Designing candidate gene and genome-wide case-control association studies," *Nat. Protoc.*, vol. 2, no. 10, pp. 2492–2501, 2007, doi: 10.1038/nprot.2007.366.
[54] J. Lasky-Su, "Chapter 19 - Statistical Techniques for Genetic Analysis," in *Clinical and Translational Science (Second Edition)*, D. Robertson and G. H. Williams, Eds. Academic Press, 2017, pp. 347–362. doi: 10.1016/B978-0-12-802101-9.00019-3.
[55] R. M. Cantor, K. Lange, and J. S. Sinsheimer, "Prioritizing GWAS results: A review of statistical methods and recommendations for their application," *Am. J. Hum. Genet.*, vol. 86, no. 1, pp. 6–22, Jan. 2010, doi: 10.1016/j.ajhg.2009.11.017.
[56] K. Kapur, "Chapter 14 - Principles of Biostatistics," in *Clinical and Translational Science (Second Edition)*, D. Robertson and G. H. Williams, Eds. Academic Press, 2017, pp. 243–260. doi: 10.1016/B978-0-12-802101-9.00014-4.
[57] S. Purcell *et al.*, "PLINK: a tool set for whole-genome association and population-based linkage analyses," *Am. J. Hum. Genet.*, vol. 81, no. 3, pp. 559–575, Sep. 2007, doi: 10.1086/519795.
[58] W. Zhou *et al.*, "Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies," *Nat. Genet.*, vol. 50, no. 9, pp. 1335–1341, Sep. 2018, doi: 10.1038/s41588-018-0184-y.
[59] V. Tam, N. Patel, M. Turcotte, Y. Bossé, G. Paré, and D. Meyre, "Benefits and limitations of genome-wide association studies," *Nat. Rev. Genet.*, vol. 20, no. 8, pp. 467–484, Aug. 2019, doi: 10.1038/s41576-019-0127-1.
[60] R. Madelaine *et al.*, "A screen for deeply conserved non-coding GWAS SNPs uncovers a MIR-9-2 functional mutation associated to retinal vasculature defects in human," *Nucleic Acids Res.*, vol. 46, no. 7, pp. 3517–3531, Apr. 2018, doi: 10.1093/nar/gky166.
[61] Z. Qu and D. L. Adelson, "Evolutionary conservation and functional roles of ncRNA," *Front. Genet.*, vol. 3, p. 205, Oct. 2012, doi: 10.3389/fgene.2012.00205.
[62] K. Watanabe, E. Taskesen, A. van Bochoven, and D. Posthuma, "Functional mapping and annotation of genetic associations with FUMA," *Nat. Commun.*, vol. 8, no. 1, p. 1826, Nov. 2017, doi: 10.1038/s41467-017-01261-5.
[63] C. A. de Leeuw, J. M. Mooij, T. Heskes, and D. Posthuma, "MAGMA: generalized gene-set analysis of GWAS data," *PLoS Comput. Biol.*, vol. 11, no. 4, p. e1004219, Apr. 2015, doi: 10.1371/journal.pcbi.1004219.
[64] F. W. Albert and L. Kruglyak, "The role of regulatory variation in complex traits and disease," *Nat. Rev. Genet.*, vol. 16, no. 4, pp. 197–212, Apr. 2015, doi: 10.1038/nrg3891.

[65] M. T. Maurano *et al.*, “Systematic localization of common disease-associated variation in regulatory DNA,” *Science*, vol. 337, no. 6099, pp. 1190–1195, Sep. 2012, doi: 10.1126/science.1222794.
[66] J. E. Grisel and J. C. Crabbe, “Quantitative Trait Loci Mapping,” *Alcohol Health Res. World*, vol. 19, no. 3, pp. 220–227, 1995, [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/31798043
[67] R. Breitling *et al.*, “Genetical genomics: spotlight on QTL hotspots,” *PLoS Genet.*, vol. 4, no. 10, p. e1000232, Oct. 2008, doi: 10.1371/journal.pgen.1000232.
[68] GTEx Consortium *et al.*, “Genetic effects on gene expression across human tissues,” *Nature*, vol. 550, p. 204, Oct. 2017, doi: 10.1038/nature24277.
[69] M. Kasowski *et al.*, “Variation in Transcription Factor Binding Among Humans,” *Science*, vol. 328, no. 5975, pp. 232–235, Apr. 2010, doi: 10.1126/science.1183621.
[70] B. E. Stranger *et al.*, “Patterns of cis regulatory variation in diverse human populations,” *PLoS Genet.*, vol. 8, no. 4, p. e1002639, Apr. 2012, doi: 10.1371/journal.pgen.1002639.
[71] H.-J. Westra and L. Franke, “From genome to function by studying eQTLs,” *Biochim. Biophys. Acta*, vol. 1842, no. 10, pp. 1896–1902, Oct. 2014, doi: 10.1016/j.bbadis.2014.04.024.
[72] M. V. Rockman and L. Kruglyak, “Genetics of global gene expression,” *Nat. Rev. Genet.*, vol. 7, no. 11, pp. 862–872, Nov. 2006, doi: 10.1038/nrg1964.
[73] N. Shan, Z. Wang, and L. Hou, “Identification of trans-eQTLs using mediation analysis with multiple mediators,” *BMC Bioinformatics*, vol. 20, no. Suppl 3, p. 126, Mar. 2019, doi: 10.1186/s12859-019-2651-6.
[74] B. L. Pierce *et al.*, “Mediation analysis demonstrates that trans-eQTLs are often explained by cis-mediation: a genome-wide analysis among 1,800 South Asians,” *PLoS Genet.*, vol. 10, no. 12, p. e1004818, Dec. 2014, doi: 10.1371/journal.pgen.1004818.
[75] B. E. Stranger *et al.*, “Population genomics of human gene expression,” *Nat. Genet.*, vol. 39, no. 10, pp. 1217–1224, Oct. 2007, doi: 10.1038/ng2142.
[76] Y. Ye, Z. Zhang, Y. Liu, L. Diao, and L. Han, “A Multi-Omics Perspective of Quantitative Trait Loci in Precision Medicine,” *Trends Genet.*, vol. 36, no. 5, pp. 318–336, May 2020, doi: 10.1016/j.tig.2020.01.009.
[77] O. Delaneau, H. Ongen, A. A. Brown, A. Fort, N. I. Panousis, and E. T. Dermitzakis, “A complete tool set for molecular QTL discovery and analysis,” *Nat. Commun.*, vol. 8, p. 15452, May 2017, doi: 10.1038/ncomms15452.
[78] A. A. Shabalin, “Matrix eQTL: ultra fast eQTL analysis via large matrix operations,” *Bioinformatics*, vol. 28, no. 10, pp. 1353–1358, May 2012, doi: 10.1093/bioinformatics/bts163.
[79] X. Zhou and M. Stephens, “Genome-wide efficient mixed-model analysis for association studies,” *Nat. Genet.*, vol. 44, no. 7, pp. 821–824, Jun. 2012, doi: 10.1038/ng.2310.
[80] A. Lewin *et al.*, “MT-HESS: an efficient Bayesian approach for simultaneous association detection in OMICS datasets, with application to eQTL mapping in multiple tissues,” *Bioinformatics*, vol. 32, no. 4, pp. 523–532, Feb. 2016, doi: 10.1093/bioinformatics/btv568.
[81] G. Li, D. Jima, F. A. Wright, and A. B. Nobel, “HT-eQTL: integrative expression

quantitative trait loci analysis in a large number of human tissues," *BMC Bioinformatics*, vol. 19, no. 1, p. 95, Mar. 2018, doi: 10.1186/s12859-018-2088-3.
[82] G. C. Imholte, M.-P. Scott-Boyer, A. Labbe, C. F. Deschepper, and R. Gottardo, "iBMQ: a R/Bioconductor package for integrated Bayesian modeling of eQTL data," *Bioinformatics*, vol. 29, no. 21, pp. 2797–2798, Nov. 2013, doi: 10.1093/bioinformatics/btt485.
[83] G. R. Abecasis, S. S. Cherny, W. O. Cookson, and L. R. Cardon, "Merlin--rapid analysis of dense genetic maps using sparse gene flow trees," *Nat. Genet.*, vol. 30, no. 1, pp. 97–101, Jan. 2002, doi: 10.1038/ng786.
[84] S. Beretta, M. Castelli, I. Gonçalves, I. Kel, V. Giansanti, and I. Merelli, "Improving eQTL Analysis Using a Machine Learning Approach for Data Integration: A Logistic Model Tree Solution," *J. Comput. Biol.*, vol. 25, no. 10, pp. 1091–1105, Oct. 2018, doi: 10.1089/cmb.2017.0167.
[85] "R/qtl software for mapping quantitative trait loci." http://rqtl.org (accessed Mar. 26, 2021).
[86] R. J. Pruim *et al.*, "LocusZoom: regional visualization of genome-wide association scan results," *Bioinformatics*, vol. 26, no. 18, pp. 2336–2337, Sep. 2010, doi: 10.1093/bioinformatics/btq419.
[87] "eMap." http://www.bios.unc.edu/~weisun/software.htm (accessed Mar. 26, 2021).
[88] H. Ongen, A. Buil, A. A. Brown, E. T. Dermitzakis, and O. Delaneau, "Fast and efficient QTL mapper for thousands of molecular phenotypes," *Bioinformatics*, vol. 32, no. 10, pp. 1479–1485, May 2016, doi: 10.1093/bioinformatics/btv722.
[89] Q. Q. Huang, S. C. Ritchie, M. Brozynska, and M. Inouye, "Power, false discovery rate and Winner's Curse in eQTL studies," *Nucleic Acids Res.*, vol. 46, no. 22, p. e133, Dec. 2018, doi: 10.1093/nar/gky780.
[90] "R/qtl software for mapping quantitative trait loci." http://rqtl.org (accessed Mar. 26, 2021).
[91] GTEx Consortium, "The GTEx Consortium atlas of genetic regulatory effects across human tissues," *Science*, vol. 369, no. 6509, pp. 1318–1330, Sep. 2020, doi: 10.1126/science.aaz1776.
[92] D. J. Burgess, "Reaching completion for GTEx," *Nature reviews. Genetics*, vol. 21, no. 12. p. 717, Dec. 2020. doi: 10.1038/s41576-020-00296-7.
[93] "eQTL Catalogue – a new resource for gene expression and splicing QTLs." https://www.ebi.ac.uk/about/news/press-releases/eQTL-catalogue (accessed Jul. 21, 2020).
[94] R. Ratnapriya *et al.*, "Retinal transcriptome and eQTL analyses identify genes associated with age-related macular degeneration," *Nat. Genet.*, vol. 51, no. 4, pp. 606–610, Apr. 2019, doi: 10.1038/s41588-019-0351-9.
[95] H. Li and H. Deng, "Systems genetics, bioinformatics and eQTL mapping," *Genetica*, vol. 138, no. 9–10, pp. 915–924, Oct. 2010, doi: 10.1007/s10709-010-9480-x.
[96] M. Wang and S. Xu, "Statistical power in genome-wide association studies and quantitative trait locus mapping," *Heredity* , vol. 123, no. 3, pp. 287–306, Sep. 2019, doi: 10.1038/s41437-019-0205-3.
[97] P. C. Sham and S. M. Purcell, "Statistical power and significance testing in large-scale genetic studies," *Nat. Rev. Genet.*, vol. 15, no. 5, pp. 335–346, May 2014, doi: 10.1038/nrg3706.

[98] X. Dong, X. Li, T.-W. Chang, S. T. Weiss, and W. Qiu, "powerEQTL: An R package and shiny application for sample size and power calculation of bulk tissue and single-cell eQTL analysis," *Cold Spring Harbor Laboratory*, p. 2020.12.15.422954, Dec. 16, 2020. doi: 10.1101/2020.12.15.422954.
[99] C. Giambartolomei *et al.*, "Bayesian test for colocalisation between pairs of genetic association studies using summary statistics," *PLoS Genet.*, vol. 10, no. 5, p. e1004383, May 2014, doi: 10.1371/journal.pgen.1004383.
[100] Y. Deng and W. Pan, "A powerful and versatile colocalization test," *PLoS Comput. Biol.*, vol. 16, no. 4, p. e1007778, Apr. 2020, doi: 10.1371/journal.pcbi.1007778.
[101] F. Hormozdiari *et al.*, "Colocalization of GWAS and eQTL Signals Detects Target Genes," *Am. J. Hum. Genet.*, vol. 99, no. 6, pp. 1245–1260, Dec. 2016, doi: 10.1016/j.ajhg.2016.10.003.
[102] S. K. Sieberts *et al.*, "Large eQTL meta-analysis reveals differing patterns between cerebral cortical and cerebellar brain regions," *Sci Data*, vol. 7, no. 1, p. 340, Oct. 2020, doi: 10.1038/s41597-020-00642-8.
[103] A. Gusev *et al.*, "Integrative approaches for large-scale transcriptome-wide association studies," *Nat. Genet.*, vol. 48, no. 3, pp. 245–252, Mar. 2016, doi: 10.1038/ng.3506.
[104] "TWAS / FUSION." http://gusevlab.org/projects/fusion/ (accessed Aug. 27, 2020).
[105] E. B. Bookman *et al.*, "Gene-environment interplay in common complex diseases: forging an integrative model—recommendations from an NIH workshop," *Genet. Epidemiol.*, vol. 35, no. 4, pp. 217–225, May 2011, doi: 10.1002/gepi.20571.
[106] "Age-Related Macular Degeneration: Facts & Figures," Jul. 04, 2015. https://www.brightfocus.org/macular/article/age-related-macular-facts-figures (accessed Apr. 20, 2021).
[107] S. Keel, J. Xie, J. Foreman, P. van Wijngaarden, H. R. Taylor, and M. Dirani, "Prevalence of Age-Related Macular Degeneration in Australia: The Australian National Eye Health Survey," *JAMA Ophthalmol.*, vol. 135, no. 11, pp. 1242–1249, Nov. 2017, doi: 10.1001/jamaophthalmol.2017.4182.
[108] Genetics Home Reference, "Age-related macular degeneration," *Genetics Home Reference*. https://ghr.nlm.nih.gov/condition/age-related-macular-degeneration (accessed Jan. 23, 2020).
[109] W. L. Wong *et al.*, "Global prevalence of age-related macular degeneration and disease burden projection for 2020 and 2040: a systematic review and meta-analysis," *Lancet Glob Health*, vol. 2, no. 2, pp. e106–16, Feb. 2014, doi: 10.1016/S2214-109X(13)70145-1.
[110] Z. Wu, L. N. Ayton, C. D. Luu, P. N. Baird, and R. H. Guymer, "Reticular Pseudodrusen in Intermediate Age-Related Macular Degeneration: Prevalence, Detection, Clinical, Environmental, and Genetic Associations," *Invest. Ophthalmol. Vis. Sci.*, vol. 57, no. 3, pp. 1310–1316, Mar. 2016, doi: 10.1167/iovs.15-18682.
[111] R. Guymer, "The genetics of age-related macular degeneration," *Clin. Exp. Optom.*, vol. 84, no. 4, pp. 182–189, Jul. 2001, doi: 10.1111/j.1444-0938.2001.tb05023.x.
[112] R. R. Priya, E. Y. Chew, and A. Swaroop, "Genetic studies of age-related macular degeneration: lessons, challenges, and opportunities for disease management," *Ophthalmology*, vol. 119, no. 12, pp. 2526–2536, Dec. 2012, doi: 10.1016/j.ophtha.2012.06.042.

[113] M. Fleckenstein *et al.*, "Age-related macular degeneration," *Nat Rev Dis Primers*, vol. 7, no. 1, p. 31, May 2021, doi: 10.1038/s41572-021-00265-2.
[114] T. Sepp *et al.*, "Complement factor H variant Y402H is a major risk determinant for geographic atrophy and choroidal neovascularization in smokers and nonsmokers," *Invest. Ophthalmol. Vis. Sci.*, vol. 47, no. 2, pp. 536–540, Feb. 2006, doi: 10.1167/iovs.05-1143.
[115] X. Han, P. Gharahkhani, P. Mitchell, G. Liew, A. W. Hewitt, and S. MacGregor, "Genome-wide meta-analysis identifies novel loci associated with age-related macular degeneration," *J. Hum. Genet.*, vol. 65, no. 8, pp. 657–665, Aug. 2020, doi: 10.1038/s10038-020-0750-x.
[116] R. Cascella *et al.*, "Towards the application of precision medicine in Age-Related Macular Degeneration," *Prog. Retin. Eye Res.*, vol. 63, pp. 132–146, Mar. 2018, doi: 10.1016/j.preteyeres.2017.11.004.
[117] F. Blond and T. Léveillard, "Functional Genomics of the Retina to Elucidate its Construction and Deconstruction," *Int. J. Mol. Sci.*, vol. 20, no. 19, Oct. 2019, doi: 10.3390/ijms20194922.
[118] M. Li *et al.*, "CFH haplotypes without the Y402H coding variant show strong association with susceptibility to age-related macular degeneration," *Nat. Genet.*, vol. 38, no. 9, pp. 1049–1054, Sep. 2006, doi: 10.1038/ng1871.
[119] M. Landowski *et al.*, "Human complement factor H Y402H polymorphism causes an age-related macular degeneration phenotype and lipoprotein dysregulation in mice," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 116, no. 9, pp. 3703–3711, Feb. 2019, doi: 10.1073/pnas.1814014116.
[120] J.-S. Joyal, M. L. Gantner, and L. E. H. Smith, "Retinal energy demands control vascular supply of the retina in development and disease: The role of neuronal lipid and glucose metabolism," *Prog. Retin. Eye Res.*, vol. 64, pp. 131–156, May 2018, doi: 10.1016/j.preteyeres.2017.11.002.
[121] Z. Yang *et al.*, "A variant of the HTRA1 gene increases susceptibility to age-related macular degeneration," *Science*, vol. 314, no. 5801, pp. 992–993, Nov. 2006, doi: 10.1126/science.1133811.
[122] M. M. DeAngelis *et al.*, "Genetics of age-related macular degeneration (AMD)," *Hum. Mol. Genet.*, vol. 26, no. R2, p. R246, Oct. 2017, doi: 10.1093/hmg/ddx343.
[123] A. R. Waksmunski, M. Grunin, T. G. Kinzy, R. P. Igo Jr, J. L. Haines, and J. N. Cooke Bailey, "Statistical driver genes as a means to uncover missing heritability for age-related macular degeneration," *BMC Med. Genomics*, vol. 13, no. 1, p. 95, Jul. 2020, doi: 10.1186/s12920-020-00747-4.
[124] L. G. Fritsche *et al.*, "A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants," *Nat. Genet.*, vol. 48, no. 2, pp. 134–143, Feb. 2016, doi: 10.1038/ng.3448.
[125] D. A. Ferrington, C. R. Fisher, and R. A. Kowluru, "Mitochondrial Defects Drive Degenerative Retinal Diseases," *Trends Mol. Med.*, vol. 26, no. 1, pp. 105–118, Jan. 2020, doi: 10.1016/j.molmed.2019.10.008.
[126] L. G. Fritsche, R. N. Fariss, D. Stambolian, G. R. Abecasis, C. A. Curcio, and A. Swaroop, "Age-related macular degeneration: genetics and biology coming together," *Annu. Rev. Genomics Hum. Genet.*, vol. 15, pp. 151–171, Apr. 2014, doi: 10.1146/annurev-genom-090413-025610.

[127] T. S. Scerri *et al.*, "Genome-wide analyses identify common variants associated with macular telangiectasia type 2," *Nat. Genet.*, vol. 49, no. 4, pp. 559–567, Apr. 2017, doi: 10.1038/ng.3799.
[128] "Macular telangiectasia type 2 | Genetic and Rare Diseases Information Center (GARD) – an NCATS Program." https://rarediseases.info.nih.gov/diseases/10690/macular-telangiectasia-type-2 (accessed Jan. 16, 2020).
[129] T. E. Clemons *et al.*, "Baseline characteristics of participants in the natural history study of macular telangiectasia (MacTel) MacTel Project Report No. 2," *Ophthalmic Epidemiol.*, vol. 17, no. 1, pp. 66–73, Jan. 2010, doi: 10.3109/09286580903450361.
[130] R. F. Spaide, "Macular Telangiectasis Type 2," in *Macular Disorders*, I. K. Kim, Ed. Singapore: Springer Singapore, 2020, pp. 73–84. doi: 10.1007/978-981-15-3001-2_8.
[131] P. Charbel Issa *et al.*, "Macular telangiectasia type 2," *Prog. Retin. Eye Res.*, vol. 34, pp. 49–77, May 2013, doi: 10.1016/j.preteyeres.2012.11.002.
[132] R. Bonelli *et al.*, "Identification of genetic factors influencing metabolic dysregulation and retinal support for MacTel, a retinal disorder," *Commun Biol*, vol. 4, no. 1, p. 274, Mar. 2021, doi: 10.1038/s42003-021-01788-w.
[133] M. L. Gantner *et al.*, "Serine and Lipid Metabolism in Macular Disease and Peripheral Neuropathy," *N. Engl. J. Med.*, vol. 381, no. 15, pp. 1422–1433, Oct. 2019, doi: 10.1056/NEJMoa1815111.
[134] R. Bonelli *et al.*, "Systemic lipid dysregulation is a risk factor for macular neurodegenerative disease," *Sci. Rep.*, vol. 10, no. 1, p. 12165, Jul. 2020, doi: 10.1038/s41598-020-69164-y.
[135] M. Pinelli *et al.*, "An atlas of gene expression and gene co-regulation in the human retina," *Nucleic Acids Res.*, vol. 44, no. 12, pp. 5773–5784, Jul. 2016, doi: 10.1093/nar/gkw486.
[136] T. Strunz *et al.*, "A mega-analysis of expression quantitative trait loci in retinal tissue," *PLoS Genet.*, vol. 16, no. 9, p. e1008934, Sep. 2020, doi: 10.1371/journal.pgen.1008934.
[137] C. A. Anderson, F. H. Pettersson, G. M. Clarke, L. R. Cardon, A. P. Morris, and K. T. Zondervan, "Data quality control in genetic case-control association studies," *Nat. Protoc.*, vol. 5, no. 9, pp. 1564–1573, Sep. 2010, doi: 10.1038/nprot.2010.116.
[138] P.-R. Loh, P. F. Palamara, and A. L. Price, "Fast and accurate long-range phasing in a UK Biobank cohort," *Nat. Genet.*, vol. 48, no. 7, pp. 811–816, Jul. 2016, doi: 10.1038/ng.3571.
[139] "Minimac4 - Genome Analysis Wiki." https://genome.sph.umich.edu/wiki/Minimac4 (accessed Jun. 24, 2021).
[140] C. M. Lee *et al.*, "UCSC Genome Browser enters 20th year," *Nucleic Acids Res.*, vol. 48, no. D1, pp. D756–D761, Jan. 2020, doi: 10.1093/nar/gkz1012.
[141] Bioconductor Package Maintainer, "liftOver: Changing genomic coordinate systems with rtracklayer::liftOver." 2019. [Online]. Available: https://www.bioconductor.org/help/workflows/liftOver/
[142] A. R. Quinlan and I. M. Hall, "BEDTools: a flexible suite of utilities for comparing genomic features," *Bioinformatics*, vol. 26, no. 6, pp. 841–842, Mar. 2010, doi: 10.1093/bioinformatics/btq033.

[143] A. M. Bolger, M. Lohse, and B. Usadel, “Trimmomatic: a flexible trimmer for Illumina sequence data,” *Bioinformatics*, vol. 30, no. 15, pp. 2114–2120, Aug. 2014, doi: 10.1093/bioinformatics/btu170.
[144] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, “edgeR: a Bioconductor package for differential expression analysis of digital gene expression data,” *Bioinformatics*, vol. 26, no. 1, pp. 139–140, Jan. 2010, doi: 10.1093/bioinformatics/btp616.
[145] G. A. Van der Auwera *et al.*, “From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline,” *Curr. Protoc. Bioinformatics*, vol. 43, pp. 11.10.1–11.10.33, 2013, doi: 10.1002/0471250953.bi1110s43.
[146] “Picard.” http://broadinstitute.github.io/picard/ (accessed Jun. 24, 2021).
[147] GTEx Consortium, “Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans,” *Science*, vol. 348, no. 6235, pp. 648–660, May 2015, doi: 10.1126/science.1262110.
[148] “Ensembl-release-97.” http://ftp.ensembl.org/pub/release-97/fasta/homo_sapiens/dna/ (accessed Jun. 12, 2021).
[149] J. D. Storey and R. Tibshirani, “Statistical significance for genomewide studies,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 100, no. 16, pp. 9440–9445, Aug. 2003, doi: 10.1073/pnas.1530509100.
[150] W. Revelle, “Procedures for Psychological, Psychometric, and Personality Research [R package psych version 2.1.6],” Jun. 2021, Accessed: Aug. 08, 2021. [Online]. Available: https://CRAN.R-project.org/package=psych
[151] X. Dong, X. Li, T.-W. Chang, C. R. Scherzer, S. T. Weiss, and W. Qiu, “powerEQTL: An R package and shiny application for sample size and power calculation of bulk tissue and single-cell eQTL analysis,” *Bioinformatics*, May 2021, doi: 10.1093/bioinformatics/btab385.
[152] H. Song *et al.*, “Long non-coding RNA LINC01535 promotes cervical cancer progression via targeting the miR-214/EZH2 feedback loop,” *J. Cell. Mol. Med.*, vol. 23, no. 9, pp. 6098–6111, Sep. 2019, doi: 10.1111/jcmm.14476.
[153] Y.-A. Moon and J. D. Horton, “Identification of two mammalian reductases involved in the two-carbon fatty acyl elongation cascade,” *J. Biol. Chem.*, vol. 278, no. 9, pp. 7335–7343, Feb. 2003, doi: 10.1074/jbc.M211684200.
[154] B. Mohamed *et al.*, “Very-long-chain fatty acid metabolic capacity of 17-beta-hydroxysteroid dehydrogenase type 12 (HSD17B12) promotes replication of hepatitis C virus and related flaviviruses,” *Sci. Rep.*, vol. 10, no. 1, p. 4040, Mar. 2020, doi: 10.1038/s41598-020-61051-w.
[155] T. Schwarz *et al.*, “1 Powerful eQTL mapping through low coverage RNA sequencing”, doi: 10.1101/2021.08.08.455466.
[156] L. D. Orozco *et al.*, “Integration of eQTL and a Single-Cell Atlas in the Human Eye Identifies Causal Genes for Age-Related Macular Degeneration,” *Cell Rep.*, vol. 30, no. 4, pp. 1246–1259.e6, Jan. 2020, doi: 10.1016/j.celrep.2019.12.082.
[157] R. Bonelli *et al.*, “Genetic disruption of serine biosynthesis is a key driver of macular telangiectasia type 2 aetiology and progression,” *Genome Med.*, vol. 13, no. 1, p. 39, Mar. 2021, doi: 10.1186/s13073-021-00848-4.
[158] M. K. Ikram *et al.*, “Four novel Loci (19q13, 6q24, 12q24, and 5q14) influence the microcirculation in vivo,” *PLoS Genet.*, vol. 6, no. 10, p. e1001184, Oct. 2010, doi:

10.1371/journal.pgen.1001184.
[159] X. Sim *et al.*, "Genetic loci for retinal arteriolar microcirculation," *PLoS One*, vol. 8, no. 6, p. e65804, Jun. 2013, doi: 10.1371/journal.pone.0065804.
[160] S.-Y. Shin *et al.*, "An atlas of genetic influences on human blood metabolites," *Nat. Genet.*, vol. 46, no. 6, pp. 543–550, Jun. 2014, doi: 10.1038/ng.2982.
[161] B. Liu, M. J. Gloudemans, A. S. Rao, E. Ingelsson, and S. B. Montgomery, "Abundant associations with gene expression complicate GWAS follow-up," *Nat. Genet.*, vol. 51, no. 5, pp. 768–769, May 2019, doi: 10.1038/s41588-019-0404-0.
[162] R. Durbin, "Efficient haplotype matching and storage using the positional Burrows-Wheeler transform (PBWT)," *Bioinformatics*, vol. 30, no. 9, pp. 1266–1272, May 2014, doi: 10.1093/bioinformatics/btu014.
[163] M. E. Ritchie *et al.*, "limma powers differential expression analyses for RNA-sequencing and microarray studies," *Nucleic Acids Res.*, vol. 43, no. 7, p. e47, Apr. 2015, doi: 10.1093/nar/gkv007.
[164] G. L. Lehmann, I. Benedicto, N. J. Philp, and E. Rodriguez-Boulan, "Plasma membrane protein polarity and trafficking in RPE cells: past, present and future," *Exp. Eye Res.*, vol. 126, pp. 5–15, Sep. 2014, doi: 10.1016/j.exer.2014.04.021.
[165] R. Simó, M. Villarroel, L. Corraliza, C. Hernández, and M. Garcia-Ramírez, "The retinal pigment epithelium: something more than a constituent of the blood-retinal barrier--implications for the pathogenesis of diabetic retinopathy," *J. Biomed. Biotechnol.*, vol. 2010, p. 190724, Feb. 2010, doi: 10.1155/2010/190724.
[166] T. Sinha, M. I. Naash, and M. R. Al-Ubaidi, "The Symbiotic Relationship between the Neural Retina and Retinal Pigment Epithelium Is Supported by Utilizing Differential Metabolic Pathways," *iScience*, vol. 23, no. 4, p. 101004, Apr. 2020, doi: 10.1016/j.isci.2020.101004.
[167] K. Eade *et al.*, "Serine biosynthesis defect due to haploinsufficiency of PHGDH causes retinal disease," *Nat Metab*, vol. 3, no. 3, pp. 366–377, Mar. 2021, doi: 10.1038/s42255-021-00361-3.
[168] A. May *et al.*, "Ongoing controversies and recent insights of the ARMS2-HTRA1 locus in age-related macular degeneration," *Exp. Eye Res.*, p. 108605, Apr. 2021, doi: 10.1016/j.exer.2021.108605.
[169] S. M. Urbut, G. Wang, P. Carbonetto, and M. Stephens, "Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions," *Nat. Genet.*, vol. 51, no. 1, pp. 187–195, Jan. 2019, doi: 10.1038/s41588-018-0268-8.
[170] T. Raj *et al.*, "Integrative transcriptome analyses of the aging brain implicate altered splicing in Alzheimer's disease susceptibility," *Nat. Genet.*, vol. 50, no. 11, pp. 1584–1592, Nov. 2018, doi: 10.1038/s41588-018-0238-1.